Probability with Measure

Dr Nic Freeman

February 2, 2024

Contents

Chapter 0

Introduction

0.1 Organization

0.1.1 Syllabus

These notes are for two courses: MAS31002 and MAS61022.

Some sections of the course are included in MAS61022 but not in MAS31002. These sections are marked with a (Δ) symbol. We will not cover these sections in lectures. Students taking MAS61022 should study these sections independently.

Some parts of the notes are marked with a ($\star$) symbol, which means they are off-syllabus. These are often cases where detailed connections can be made to and from other parts of mathematics.

0.1.2 Problem sheets

The exercises are divided up according to the chapters of the course. Some exercises are marked as 'challenge questions' – these are intended to offer a serious, time consuming challenge to the best students.

Aside from challenge questions, it is expected that students will attempt all exercises (for the version of the course they are taking) and review their own solutions using the typed solutions provided in the online version of these notes, in Appendix B.

At three points during each semester, an assignment of additional exercises will be set. About one week later, a mark scheme will be posted, and you should self-mark your solutions.

0.1.3 Examination

The course will be examined in the summer sitting. Parts of the course marked with a (Δ) are examinable for MAS61022 but not for MAS31002. Parts of the course marked with a ($\star$) will not be examined (for everyone). Some advice on how to structure your revision can be found in Appendix A.

0.1.4 Website

Further information, including the timetable, can be found on

`https://nicfreeman1209.github.io/Website/MASx50/`.

0.2 Preliminaries

This section contains lots of definitions, mostly from earlier courses, that we will use later on. It should be familiar to you but there may be one or two minor extensions of ideas you have seen before.

1. *Set Theory.*

   Let S be a set, with subsets A, B and A_n.

   Complement: $A^c = \{x \in S; x \notin A\}$.

   Union: $A \cup B = \{x \in S; x \in A \text{ or } x \in B\}$.

   Intersection: $A \cap B = \{x \in S; x \in A \text{ and } x \in B\}$.

   Set theoretic difference: $A \setminus B = A \cap B^c$.

   Finite unions and intersections: $\bigcup_{i=1}^n A_i = A_1 \cup A_2 \cup \ldots \cup A_n$ and $\bigcap_{i=1}^n A_i = A_1 \cap A_2 \cap \ldots \cap A_n$.

   More generally, if I is some set and $A_i \subseteq S$ for all $i \in I$ then we define

   $$\bigcup_{i \in I} A_i = \{x\,;\,x \in A_i \text{ for some } i \in I\} \qquad \bigcap_{i \in I} A_i = \{x\,;\,x \in A_i \text{ for all } i \in I\}.$$

   Countable unions and intersections are precisely the case $I = \mathbb{N}$, usually written as $\bigcup_{i=1}^\infty A_i$ and $\bigcup_{i=1}^\infty A_i$.

   De Morgan's laws state that:

   $$S \setminus \left(\bigcap_{i \in I} A_i\right) = \bigcup_{i \in I} S \setminus A_i,$$

   $$S \setminus \left(\bigcup_{i \in I} A_i\right) = \bigcap_{i \in I} S \setminus A_i.$$

   The Cartesian product of sets S and T is the set $S \times T = \{(s,t)\,;\,s \in s, t \in T\}$.

2. *Sets of Numbers*

   - Natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$.
   - Non-negative integers $\mathbb{Z}_+ = \mathbb{N} \cup \{0\} = \{0, 1, 2, 3, \ldots\}$.
   - Integers $\mathbb{Z}$.
   - Rational numbers $\mathbb{Q}$.
   - Real numbers $\mathbb{R}$.
   - Complex numbers $\mathbb{C}$.

   A set X is *countable* if there exists an injection between X and $\mathbb{N}$. A set is *uncountable* if it fails to be countable. $\mathbb{N}, \mathbb{Z}_+, \mathbb{Z}$ and $\mathbb{Q}$ are countable. $\mathbb{R}$ and $\mathbb{C}$ are uncountable. All finite sets are countable.

3. *Images and Preimages.*

   Suppose that S_1 and S_2 are two sets and that $f : S_1 \to S_2$ is a mapping (or function). Suppose that $A \subseteq S_1$. The *image* of A under f is the set $f(A) \subseteq S_2$ defined by

   $$f(A) = \{y \in S_2; y = f(x) \text{ for some } x \in S_1\}.$$

   If $B \subseteq S_2$ the *inverse image* or *pre-image* of B under f is the set $f^{-1}(B) \subseteq S_1$ defined by

   $$f^{-1}(B) = \{x \in S_1; f(x) \in B\}.$$

   Note that $f^{-1}(B)$ makes sense irrespective of whether the mapping f is invertible.

   Key properties are, with $A, A_1, A_2 \subseteq S_1$ and $B, B_1, B_2 \subseteq S_2$:

   $$\begin{aligned} f^{-1}(B_1 \cup B_2) &= f^{-1}(B_1) \cup f^{-1}(B_2), \\ f^{-1}(B_1 \cap B_2) &= f^{-1}(B_1) \cap f^{-1}(B_2), \\ f^{-1}(A^c) &= f^{-1}(A)^c, \\ f(A_1 \cup A_2) &= f(A_1) \cup f(A_2), \\ f(A_1 \cap A_2) &\subseteq f(A_1) \cap f(A_2), \end{aligned}$$

   Note also that if $A \subseteq B$ then $f(A) \subseteq f(B)$ and $f^{-1}(A) \subseteq f^{-1}(B)$.

4. *Extended Real Numbers*

   We will often find it convenient to work with ∞ and $-\infty$. These are *not* real numbers, but we find it convenient to treat them a bit like real numbers. To do so we specify some extra arithmetic rules:

   - for all $x \in \mathbb{R}$ we have $\infty + x = x + \infty = \infty$,
   - for $x > 0$ we have $x \times \infty = \infty \times x = \infty$,
   - for all $x \in \mathbb{R}$ we have $\frac{x}{\infty} = 0$ and $\frac{\infty}{x} = \infty$,
   - $\infty \times (-1) = -\infty$ and $(-\infty) \times (-1) = \infty$.

   Combining these rules and using the usual properties of real arithmetic (e.g. $a \times b = b \times a$) allows us to deduce further properties, for example for $x < 0$ we have $x \times \infty = (-1) \times (-x) \times \infty = (-1) \times \infty = -\infty$. Any arithmetic expressions involving $\pm\infty$ that are not specified by the above rules are undefined. In particular, $\infty - \infty$, $0 \times \infty$ and $\frac{\infty}{\infty}$ are undefined.

   We write $\overline{\mathbb{R}} = \{-\infty\} \cup \mathbb{R} \cup \{\infty\}$, which is known as the *extended* real numbers. We also specify that, for all $x \in \mathbb{R}$,

   $$-\infty < x < \infty.$$

5. *Analysis.*

   - sup and inf. If A is a bounded set of real numbers, we write $\sup(A)$ and $\inf(A)$ for the real numbers that are their least upper bounds and greatest lower bounds (respectively.) If A fails to be bounded above, we write $\sup(A) = \infty$ and if A fails to be bounded below we write $\inf(A) = -\infty$. Note that $\inf(A) = -\sup(-A)$ where $-A = \{-x\,;\, x \in A\}$. If $f : S \to \mathbb{R}$ is a mapping, we write $\sup_{x\in S} f(x) = \sup\{f(x); x \in S\}$. A very useful inequality is

     $$\sup_{x\in S} |f(x) + g(x)| \leq \sup_{x\in S} |f(x)| + \sup_{x\in S} |g(x)|.$$

- Sequences and Limits. Let $(a_n) = (a_1, a_2, a_3, \ldots)$ be a sequence of real numbers. It *converges* to the real number a if given any $\epsilon > 0$ there exists a natural number N so that whenever $n > N$ we have $|a - a_n| < \epsilon$. We then write $a = \lim_{n\to\infty} a_n$.

  A sequence (a_n) which is *monotonic increasing* (i.e. $a_n \leq a_{n+1}$ for all $n \in \mathbb{N}$) and *bounded above* (i.e. there exists $K > 0$ so that $a_n \leq K$ for all $n \in \mathbb{N}$) converges to $\sup_{n\in\mathbb{N}} a_n$.

  A sequence (a_n) which is *monotonic decreasing* (i.e. $a_{n+1} \leq a_n$ for all $n \in \mathbb{N}$) and *bounded below* (i.e. there exists $L > 0$ so that $a_n \geq L$ for all $n \in \mathbb{N}$) converges to $\inf_{n\in\mathbb{N}} a_n$.

  A *subsequence* of a sequence (a_n) is itself a sequence of the form (a_{r_n}) where $r_n < r_{n+1}$ for all $n \in \mathbb{N}$.

- Series. If the sequence (s_n) converges to a limit s where $s_n = a_1 + a_2 + \cdots + a_n$ we write $s = \sum_{n=1}^{\infty} a_n$ and call it the *sum of the series*. If each $a_n \geq 0$ then the sequence (s_n) is either convergent to a limit or properly divergent to infinity. In the latter case we write $s = \infty$ and interpret this in the sense of extended real numbers.

- Continuity. A function $f : \mathbb{R} \to \mathbb{R}$ is *continuous* at $a \in \mathbb{R}$ if given any $\epsilon > 0$ there exists $\delta > 0$ so that $|x - a| < \delta \Rightarrow |f(x) - f(a)| < \epsilon$. Equivalently f is continuous at a if given any sequence (a_n) that converges to a, the sequence $(f(a_n))$ converges to $f(a)$.

  f is a *continuous function* if it is continuous at every $a \in \mathbb{R}$.

Chapter 1

Measure Spaces

1.1 What is measure theory?

Measure theory is the abstract mathematical theory that underlies all models of measurement of 'size' in the real world. This includes measurement of length, area and volume, weight and mass, and also of chance and probability. Measure theory is a branch of pure mathematics, in particular of analysis, but it plays key roles in both calculus and statistical modelling. This is because measure theory provides the foundation of both the modern theory of integration and of the modern theory of probability.

Suppose that we wish to measure the lengths of several line segments. We represent these as closed intervals of the real number line $\mathbb{R}$ so a typical line segment is $[a,b]$ where $b > a$. We all agree that its length is $b - a$. We write this as

$$m([a,b]) = b - a$$

and interpret this as telling us that the measure m of length of the line segment $[a,b]$ is the number $b - a$. We might also agree that if $[a_1,b_1]$ and $[a_2,b_2]$ are two non-overlapping line segments and we want to measure their combined length then we want to apply m to the set-theoretic union $[a_1,b_1] \cup [a_2,b_2]$ and

$$m([a_1,b_1] \cup [a_2,b_2]) \quad = (b_2 - a_2) + (b_1 - a_1) = m([a_1,b_1]) + m([a_2,b_2]). \tag{1.1}$$

An isolated point c has zero length and so

$$m(\{c\}) = 0.$$

If we consider the whole real line in its entirety then it has infinite length, i.e.

$$m(\mathbb{R}) = \infty.$$

The key point here is that, if we try to abstract the notion of a 'measure of length, then we should regard it as a mapping m defined on subsets of the real line, that takes values in the extended non-negative real numbers $[0,\infty]$.

We might wonder why there is any mathematical difficulty involved here, since it appears that we can easily agree on how how long a line is. The problem is that subsets of $\mathbb{R}$ may arise naturally and still be rather complicated.

Example 1.1.1 (The Cantor Set) Let $C_0 = [0,1]$. Given C_n, define C_{n+1} by taking each sub-interval of C_n, cutting this sub-interval into three parts of equal length and removing the open interval corresponding to the middle third. So, for each n, C_n is a set of 2^n closed intervals each of length $(\frac{1}{3})^n$.

Let $C = \bigcap_{n=0}^{\infty} C_n$. Clearly $C_{n+1} \subseteq C_n$, so this is a decreasing sequence of sets, and C is precisely the points that 'never end up in the middle thirds'. For example, $0 \in C_n$ and $\frac{1}{3} \in C_n$.

The total length of the intervals in C_n is $2^n(\frac{1}{3})^n = (\frac{2}{3})^n$, which tends to zero as $n \to \infty$. This suggests C should have 'length' zero, but how can we make this intuition into rigorous mathematics?

The Cantor set is a 'fractal', which is a general term for any shape with very detailed structure. It is somewhat contrived – in fact, it was first introduced precisely as a contrived example of an odd looking shape that appeared to exist within the real line, but with no obvious purpose. Today, we know that fractal-like objects appear frequently within nature, which means that we also need to deal with them within our theory of measure.

1.2 Sigma fields

We need to be more ambitious that just measuring the length of intervals of $\mathbb{R}$. More generally, we want to work with a function m such that the map $A \mapsto m(A)$ corresponds to our intuitive idea of measuring how 'big' the object A is. Length is one example of this, 'weight' and 'volume' are other examples. The function m will be known as *a measure*, and we say that m *measures* the length/volume/size/weight/etc of A.

To do this rigorously, the first question we must answer is: which objects are we going to measure? This question has a reasonably straightforward answer. We are going to take a set S, and we are going to 'measure' subsets A of the set S. Note that at this stage we don't specify *what property* of A we are going to measure. We might measure length, or volume, or some other property that might be more difficult to express in words.

However, there is a caveat. In many cases, particularly if the set S is very large (such as $\mathbb{R}$ itself, which is uncountable) we will not be able to measure the size of *every* subset of S. The reasons for this caveat are difficult, and we will come to them in Section 1.6. Instead, we do the next best thing. We specify precisely *which* subsets of S we are going to measure.

Definition 1.2.1 Let S be a set. A *σ-field* on S is a set Σ, such that each $A \in \Sigma$ is a subset of S, satisfying the following properties:

(S1) $\emptyset \in \Sigma$ and $S \in \Sigma$.

(S2) If $A \in \Sigma$ then $A^c \in \Sigma$.

(S3) If $(A_n)_{n\in\mathbb{N}}$ is a sequence of sets with $A_n \in \Sigma$ for all $n \in \mathbb{N}$ then $\bigcup_{n=1}^{\infty} A_n \in \Sigma$.

Definition 1.2.2 Given a σ-field Σ, a set $A \in \Sigma$ is said to be *measurable* with respect to Σ. We will often shorten this to 'Σ-measurable', or simply 'measurable' if the context makes clear which Σ is meant.

The purpose of (S1)-(S3) is to capture some of our intuition on what it means 'to measure'. Let us go through them carefully. The first part of (S1) says that we should be able to measure a set $\emptyset$ that is empty (and, when the time comes, we will force $\emptyset$ to have measure zero). Property (S2) is a statement that if we are going to be able to measure A, we also want to be able to measure its complement $A^c = S \setminus A$. This is very natural from a physical point of view: if you have a 1kg bag of flour and you take 450g out, then you expect to be able to measure how much flour you have left. The complement of $\emptyset$ is $\emptyset^c = S \setminus \emptyset = S$, so this means we also need to be able to measure S itself, thus leading us to the second half of (S1).

Property (S3) is a bit more subtle. Firstly note that if we can measure A and B then it is reasonable (again, think flour) to want to measure their union $A \cup B$. Similarly, if we can measure $A_1, \dots, A_n$ then it is reasonable to want to measure their union $\cup_{i=1}^{n} A_n$. However, we can't stop here. We need our theory of measure to handle infinite objects, like the interval $[0, 2]$ which contains infinitely many elements (even though it only has length 2!). For this reason we also allow countable unions, of the form $\cup_{n=1}^{\infty} \dots$ in (S3).

Remark 1.2.3 We cannot 'upgrade' to allowing *uncountable* unions in (S3). Doing so would, unfortunately, break our entire theory of measure, for a reason that we cannot easily see, yet. We will discuss this point further in Section 1.6.

Remark 1.2.4 The term *σ-algebra* is used by some books, with the same meaning as σ-field. I prefer σ-field, you may use either.

Let us briefly note a few properties of σ-fields, which build on the properties (S1)-(S3).

- We have seen in (S3) that Σ is closed under countably infinite unions, meaning that taking a countable unions of sets in Σ gives back a set in Σ. The same is true of finite unions. To see this let $A_1, \ldots, A_n \in \Sigma$ and define $A_i = \emptyset$ for $i > n$. By (S1) we have $A_i \in \Sigma$ for all $i \in \mathbb{N}$. Note that $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{n} A_i$, and thus by (S3) we have $\bigcup_{i=1}^{n} A_i \in \Sigma$.
- Σ is also closed under countably infinite intersections. To see this we can use the laws of set algebra to write $\bigcap_{i=1}^{\infty} A_i = \left(\bigcup_{i=1}^{\infty} A_i^c\right)^c$, and then apply (S2) and (S3) to the right hand side. By the same ideas as above, Σ is also closed under finite intersections.
- Σ is also closed under set theoretic differences. To see this note that $A \setminus B = A \cap B^c$, and apply (S2) along with closure under intersections to the the right hand side.

We can summarise the above properties as follows: if we have a σ-field Σ, and sets $A_1, A_2, \ldots \in \Sigma$, then applying any finite or countable number of set operations to the A_i will simply give us back another set in Σ. We call this fact 'closure under countable set operations'. We will use it repeatedly throughout the course.

Definition 1.2.5 A pair (S, Σ) where S is a set and Σ is a σ-field of subsets of S is called a *measurable space.*

Given a set S, there are typically many possible choices of Σ. The choice of Σ is determined by what it is that we want to measure.

1.2.1 Examples of σ-fields

The following examples are all σ-fields.

1. For any set S, the power set $\mathcal{P}(S)$ is a σ-field. Recall that $\mathcal{P}(S)$ is the set of all subsets of S, so (S1)-(S3) are automatically satisfied.

2. For any set S, $\Sigma = \{\emptyset, S\}$ is a σ-field, called the *trivial* σ-field.

3. If S is any set and $A \subset S$ then $\Sigma = \{\emptyset, A, A^c, S\}$ is a σ-field. Checking (S1)-(S3) in this case is left for you.

4. Similarly, if $A, B \subset S$ then $\Sigma = \{\emptyset, A, B, A \cup B, (A \cup B)^c, A \cap B, (A \cap B)^c, A \setminus B, (A \setminus B)^c, B \setminus A, (B \setminus A)^c, (A \cup B) \setminus (A \cap B), ((A \cup B) \setminus (A \cap B))^c, A \cup B^c, A^c \cup B, S\}$ is a σ-field. I suggest not checking this one.

I hope this is a convincing demonstration that we cannot hope to simply write down σ-fields, for the most part. Instead we need a tool for constructing them, without needing to write them down. This is done as follows.

Lemma 1.2.6 *Let I be any set and for each $i \in I$ let Σ_i be a σ-field on S. Then*

$$\Sigma = \bigcap_{i \in I} \Sigma_i \tag{1.2}$$

is a σ-field on S.

PROOF: We check the three conditions of Definition 1.2.1 for $\mathcal{F}$.

(S1) Since each Σ_i is a σ-field, we have $\emptyset \in \Sigma_i$. Hence $\emptyset \in \cap_i \Sigma_i$. Similarly, $S \in \Sigma$.

(S2) If $A \in \Sigma = \cap_i \mathcal{F}_i$ then $A \in \Sigma_i$ for each i. Since each Σ_i is a σ-field, $S \setminus A \in \Sigma_i$ for each i. Hence $S \setminus A \in \cap_i \Sigma_i$.

(S3) If $A_j \in \Sigma$ for all j, then $A_j \in \Sigma_i$ for all i and j. Since each Σ_i is a σ-field, $\cup_j A_j \in \Sigma_i$ for all i. Hence $\cup_j A_j \in \cap_i \Sigma_i$. ■

Example 1.2.7 Thanks to Lemma 1.2.6 we can construct a σ-field by making a statement along the lines of

> *"Let Σ be the smallest σ-field on $\mathbb{R}$ containing all the open intervals."*

By this statement we mean: let Σ be intersection of all the σ-fields on $\mathbb{R}$ that contain all of the open intervals, in the style of equation (1.2). We know that at least one σ-field exists with this property, namely $\mathcal{P}(\mathbb{R})$. Therefore Lemma 1.2.6 applies, and tells that Σ is indeed a σ-field. The σ-field resulting from this example is very special. It is known as the Borel σ-field on $\mathbb{R}$, and it is much smaller than $\mathcal{P}(\mathbb{R})$. We will introduce it formally in Definition 1.4.2, and study it in Section 1.4.

1.3 Measure

The next question we need to ask is: what does it mean to measure an object? We want a general framework that we can use for concepts such as length, weight and volume. From the last section, we know that we are looking for a function $m : \Sigma \to [0,\infty]$ where Σ will be an appropriately chosen σ-field.

Definition 1.3.1 Let (S,Σ) be a measurable space. A mapping $m : \Sigma \to [0,\infty]$ is known as a *measure* if it satisfies

(M1) $m(\emptyset) = 0$.

(M2) If $(A_n)_{n\in\mathbb{N}}$ is a sequence of sets where each $A_n \in \Sigma$ and if these sets are pairwise disjoint (meaning that $A_n \cap A_m = \emptyset$ if $m \neq n$) then

$$m\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} m(A_n).$$

Note that (M2) relies on property (S3), to make sure that $\bigcup_{n=1}^{\infty} A_n \in \Sigma$. Property (M2) is often known as σ-additivity. Crucially, it will allow us to take limits in ways that involve measures, thanks to the fact that (M2) considers a (countably) infinite sequence of sets (A_n). Limits are how we rigorously justify that approximations work – consequently we need them, if we are to create a theory that will, ultimately, be useful to experimentalists and modellers.

Property (M2) encapsulates the idea that if we take a collection of objects, then their total measure should equal to the sum of their individual measures – providing they don't overlap with each other. For example, we might take 1kg of flour and divide it into 3 piles weighing 100g, 250g and 650g. We could also imagine dividing our 1kg of flour into an infinite sequence of piles, with sizes 500g, 250g, 125g, 67.5g, . . . , that sum (as an infinite series) to 1kg.

Property (M1) is much less remarkable. It simply states that the empty set has zero measure. This represents our feeling that an empty region of space has zero length/weight/volume/etc.

Definition 1.3.2 A triplet (S,Σ,m) where S is a set, Σ is a σ-field on S, and $m : \Sigma \to [0,\infty]$ is a measure is known as a *measure space.*

Definition 1.3.3 The extended real number $m(S)$ is called the *total mass* of m. The measure m is said to be *finite* if $m(S) < \infty$.

Let us now assume that (S,Σ,m) is a measure space, and record some useful properties of measures.

- If $A_1,\ldots,A_n \in \Sigma$ and are pairwise disjoint then

  $$m(A_1 \cup \ldots \cup A_n) = m(A_1) + \ldots + m(A_n).$$

  This is known as *finite additivity* of measures. We'll often think of it as part of (M2).

  To prove it we use the same idea on (M2) as we used, for σ-fields, on (S3). Define $A_i' = A_i$ for $i \leq n$ and $A_i' = \emptyset$ for $i > n$. By (M2) we have $m\left(\bigcup_{i=1}^{\infty} A_i'\right) = \sum_{i=1}^{\infty} m(A_i')$. By (M1) we have $m(\emptyset) = 0$, so this reduces to $m(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} m(A_i)$.

- If $A, B \in \Sigma$ with $A \subseteq B$ then $m(A) \leq m(B)$. This property is known as the *monotonicity* property of measures.

  To prove it write B as the disjoint union $B = (B \setminus A) \cup A$ and then use that from part 1 we have $m(B) = m((B \setminus A) \cup A) = m(B \setminus A) + m(A)$. If $m(A)$ is finite we can subtract it from both sides, and obtain that
  $$m(B \setminus A) = m(B) - m(A) \tag{1.3}$$
  However, this only works if $m(A)$ is finite!

- If $A, B \in \Sigma$ are arbitrary (i.e. not necessarily disjoint) then
  $$m(A \cup B) + m(A \cap B) = m(A) + m(B). \tag{1.4}$$
  The proof of this is Problem **1.4** part (a). Note that if $m(A \cap B) < \infty$ we have $m(A \cup B) = m(A) + m(B) - m(A \cap B)$, which you might recognize as similar to something you've seen before in probability.

1.3.1 Examples of measures

Here are three important first examples of measure spaces. We can't yet introduce examples based on length or volume; this will come later in the course.

1. **Counting Measure** Let S any set and take $\Sigma = \mathcal{P}(S)$. For each $A \subseteq S$ the counting measure $m = \#$ is given by

   $$\#(A) = \text{the number of elements in } A.$$

   I hope its intuitively obvious to you that this is a measure. We'll omit checking the details.

2. **Dirac Measure** This measure is named after the physicist Paul Dirac. Let (S, Σ) be an arbitrary measurable space and fix $x \in S$. The Dirac measure $m = \delta_x$ is defined by

   $$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

   Checking properties (M1) and (M2) in this case is left for you.

   A useful fact: if S is countable then we can write the counting measure $\#$ in terms of Dirac measures, as $\#(A) = \sum_{x \in S} \delta_x(A)$.

3. **Probability**

   Consider a finite set $S = \{x_1, \ldots, x_n\}$, which we'll call the *sample space* and call each of the x_i an *outcome*. Let Σ be the set of all subsets of S. Let $(p_i)_{i=1}^n$ be set of numbers in $[0, 1]$ such that $\sum_{i=1}^n p_i = 1$. For $A \in \Sigma$ we define a measure $m = \mathbb{P}$ by setting

   $$\mathbb{P}[A] = \sum_{i=1}^{n} p_i \delta_{x_i}(A). \tag{1.5}$$

   In words, to each outcome x_i we assign probability p_i, that is $\mathbb{P}[\{x_i\}] = p_i$. If a set A contains several outcomes, then its outcome is precisely the sum of their individual probabilities. Finding the probability of an event is just another kind of measuring!

   We could treat a countable set S similarly, with a countable sequence of p_i and a countable summation (i.e. an infinite series) in (1.5). Probability, however, mostly requires uncountable sample spaces (e.g. the normal distribution on the real line). In this case (1.5) breaks down completely, because there is no such thing as an uncountable sum. One of the outcomes of this course will be a rigorous basis for probability theory with uncountable sample spaces.

   In general, a measure m is said to be a *probability* measure if its total mass is 1 i.e. $m(S) = 1$.

4. **Integration**

   In previous analysis courses you viewed Riemann integration as a way of calculating area – that is, measuring the area of two-dimensional shapes. You've probably also viewed various types of integrals as ways of calculating volumes, at some point. So, we should expect integration to fit naturally into our theory of measures.

   In Chapter 4 we will introduce *Lebesgue integration*. Lebesgue integration is 'the' modern theory of integration on which mathematical modelling now relies. We will see that Lebesgue integration interacts nicely with measure theory, whilst Riemann integration doesn't. In fact, Lebesgue integration will also be the key tool for setting up a rigorous basis for probability theory.

1.4 The Borel σ-field

In this section introduce another example of a measure space, which will represent the notion of measuring the 'length' of subsets of $\mathbb{R}$. For an interval $[a,b]$ it is clear that the length should be $b-a$, but as we saw in Section 1.1 for more complicated subsets of $\mathbb{R}$ the situation is not so clear.

Example 1.4.1 Consider, for example, the irrational numbers $\mathbb{I}$ and the rational numbers $\mathbb{Q}$. Both $\mathbb{I}$ and $\mathbb{Q}$ are found throughout $\mathbb{R}$, but they are both full of tiny holes. Can we find a meaningful way to decide what is the 'length' of $\mathbb{I}$ and $\mathbb{Q}$? In fact, we will see that we can – to come in Section 1.5. But in Section 1.6 we will also show that it is possible to construct subsets of $\mathbb{R}$ for which there is *no* meaningful idea of length.

In this section we take $S=\mathbb{R}$. The first question is: which σ-field should we use? The power set $\mathcal{P}(\mathbb{R})$ is too big, for reasons that we will make clear in Section 1.6. However, for practical purposes we do need our σ-field to contain all open and closed intervals, and also unions, intersections and complements of these. This provides a starting point.

Definition 1.4.2 The *Borel σ-field* of $\mathbb{R}$, denoted by $\mathcal{B}(\mathbb{R})$, is the smallest σ-field on $\mathbb{R}$ that contains all open intervals (a,b) where $-\infty\leq a<b\leq\infty$. Sets in $\mathcal{B}(\mathbb{R})$ are called *Borel sets.*

Note that $\mathcal{B}(\mathbb{R})$ also contains isolated points $\{a\}$ where $a\in\mathbb{R}$. To see this first observe that $(a,\infty)\in\mathcal{B}(\mathbb{R})$ and also $(-\infty,a)\in\mathcal{B}(\mathbb{R})$. Now by (S2), $(-\infty,a]=(a,\infty)^c\in\mathcal{B}(\mathbb{R})$ and $[a,\infty)=(-\infty,a)^c\in\mathcal{B}(\mathbb{R})$. Finally as σ-fields are closed under intersections, $\{a\}=[a,\infty)\cap(-\infty,a]\in\mathcal{B}(\mathbb{R})$. You can show that $\mathcal{B}(\mathbb{R})$ also contains all closed intervals – see Problem **1.6**. With open and closed intervals in hand, the closure of σ-fields under countable set operations gives us a way to construct a huge variety of Borel sets.

As a general rule, all 'sensible' subsets of $\mathbb{R}$ are Borel sets. We might hope to find some sort of formula for a general element of $\mathcal{B}(\mathbb{R})$, but this is not possible. Unless you deliberately set out to find a non-Borel subset of $\mathbb{R}$ you will never come across one – and even when you look for them it is hard work to find them, as we will see in Section 1.6.

1.5 Lebesgue measure

The measure that precisely captures the notion of length is called *Lebesgue measure* in honour of the French mathematician Henri Lebesgue (1875-1941), who founded the modern theory of integration. We will denote it by λ. First we need a definition.

Let $A\in\mathcal{B}(\mathbb{R})$ be arbitrary. A *covering* of A is a finite or countable collection of open intervals $\{(a_n,b_n),n\in\mathbb{N}\}$ so that

$$A\subseteq\bigcup_{n=1}^{\infty}(a_n,b_n). \tag{1.6}$$

Definition 1.5.1 Let $\mathcal{C}_A$ be the set of all coverings of the set $A\in\mathcal{B}(\mathbb{R})$. The *Lebesgue measure* λ on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$ is defined by the formula:

$$\lambda(A)=\inf_{\mathcal{C}_A}\sum_{n=1}^{\infty}(b_n-a_n), \tag{1.7}$$

where the inf is taken over all possible coverings of A, with notation as in (1.6)

It would take a long time to prove that λ really is a measure, and it wouldn't help us understand λ any better if we did it, so we'll omit that from the course. You can find a proof in any standard text book on measure theory e.g. Cohn, Schilling or Tao. Instead, let's check that the Definition 1.5.1 agrees with some of our intuitive ideas about length.

(L1) If $A = (a,b)$ then $\lambda((a,b)) = b - a$ as expected, since (a,b) is a covering of itself and any other cover will have greater length.

(L2) If $A = \{a\}$ then $\lambda(\{a\}) = 0$. To see this, choose any $\epsilon > 0$. Then $(a-\epsilon/2, a+\epsilon/2)$ is a cover of a and so $\lambda(\{a\}) \le (a+\epsilon/2) - (a-\epsilon/2) = \epsilon$. But ϵ is arbitrary and so we conclude that $\lambda(\{a\}) = 0$.

(L3) Combining (L2) with (M2), we deduce that for $a < b$,

$$\lambda([a,b)) = \lambda(\{a\} \cup (a,b)) = \lambda(\{a\}) + \lambda((a,b)) = b - a.$$

Similarly, $\lambda([a,b]) = \lambda((a,b]) = b - a$.

(L4) If $A = [0,\infty)$, write $A = \bigcup_{n=1}^{\infty}[n-1,n)$. Then by (M2) we obtain $\lambda([0,\infty)) = \sum_{n=1}^{\infty} 1 = \infty$. By a similar argument, $\lambda((-\infty,0)) = \infty$ and so $\lambda(\mathbb{R}) = \lambda((-\infty,0)) + \lambda([0,\infty)) = \infty$.

(L5) If $A \in \mathcal{B}(\mathbb{R})$, and for some $x \in \mathbb{R}$ we define $A_x = \{x + a\,;\, a \in A\}$, then $\lambda(A) = \lambda(A_x)$.

In words, if we take a set A and translate it (by x), we do not change its measure. We'll often refer to this property as the *translation invariance* of Lebesgue measure. It is easily seen from (1.7), because any cover of A can be translated by x to be a cover of A_x.

Example 1.5.2 In simple practical examples on Lebesgue measure, it is best not to try to use (1.7) directly, but to just apply the properties listed above. For example, to find $\lambda\left((-3,10) \setminus (-1,4)\right)$, use (L3) and (M2) to obtain

$$\begin{aligned}\lambda\left((-3,10)\setminus(-1,4)\right) &= \lambda\left((-3,-1] \cup [4,10)\right)\\ &= \lambda((-3,-1]) + \lambda([4,10))\\ &= ((-1)-(-3)) + (10-4) = 8.\end{aligned}$$

It is possible for a set to be quite 'large' and still have Lebesgue measure zero. The next two lemmas give examples of such sets.

Lemma 1.5.3 *Let $A \subset \mathbb{R}$ be countable. Then $\lambda(A) = 0$.*

PROOF: Since A is countable we may write $A = \{a_1, a_2, \ldots\} = \bigcup_{n=1}^{\infty}\{a_n\}$. Since A is a countable union of singletons, it is in $\mathcal{B}(\mathbb{R})$. Then, using (M2) and (L2)

$$\lambda(A) = \lambda\left(\bigcup_{n=1}^{\infty}\{a_n\}\right) = \sum_{n=1}^{\infty}\lambda(\{a_n\}) = 0.$$

■

It follows that

$$\lambda(\mathbb{N}) = \lambda(\mathbb{Z}) = \lambda(\mathbb{Q}) = 0.$$

Further, for any $A \in \mathcal{B}(\mathbb{R})$ we have $\lambda(A \cap \mathbb{Q}) \le \lambda(\mathbb{Q})$, which implies $\lambda(A \cap \mathbb{Q}) = 0$. Thus also, if A has finite measure, $\lambda(A) - \lambda(A \cap \mathbb{I}) = \lambda(A \setminus (A \cap \mathbb{I})) = \lambda(A \cap \mathbb{Q}) = 0$. This is particularly intriguing as it tells us that

$$\lambda(A) = \lambda(A \cap \mathbb{I}),$$

so the only contribution to length of sets of real numbers comes from the irrational numbers. Hence also for all n, $\lambda(\mathbb{I}) \ge \lambda(\mathbb{I} \cap [-n, n]) = \lambda([-n, n]) = 2n$, and letting $n \to \infty$ gives that $\lambda(\mathbb{I}) = \infty$.

Lemma 1.5.4 *The Cantor Set has Lebesgue measure zero.*

PROOF: Recall the construction of the Cantor set $C = \bigcap_{n=1}^{\infty} C_n$ given in Example 1.1.1, and the notation used there. Recall also that the C_n are decreasing, that is $C_{n+1} \subseteq C_n$, and hence also $C \subseteq C_n$ for all n. Since C_n is a union of 2^n disjoint intervals of length 3^{-n} using (M2) and (L3) we have $\lambda(C_n) = 2^n(\frac{1}{3})^n = (\frac{2}{3})^n$. Using monotonicity of measure we thus have $0 \le \lambda(C) \le \lambda(C_n) = (\frac{2}{3})^n$. Letting $n \to \infty$, and applying the sandwich rule we obtain $\lambda(C) = 0$. ■

We'll tend to use (L1)-(L5) without explicitly referencing them, from now on. Hopefully, by this point, you're happy to trust that Lebesgue measure matches your intuitive concept of length within $\mathbb{R}$.

Remark 1.5.5 If I is a closed interval (or in fact any Borel set) in $\mathbb{R}$ we can similarly define $\mathcal{B}(I)$, the Borel σ-field of I, to be the smallest σ-field containing all open intervals in I. In fact, it holds that $\mathcal{B}(I) = \{B \cap I\,;\, B \in \mathcal{B}(\mathbb{R})\}$. The Lebesgue measure λ_I on $(I, \mathcal{B}(I))$ is obtained by restricting the sets A in (1.7) to be in $\mathcal{B}(I)$. It can be seen that for $A \subseteq I$ we have $\lambda_I(A) = \lambda(A)$. We won't include a proof of these claims.

1.6 An example of a non-measurable set $(\star)$

Note that this section has a $(\star)$, meaning that it is off-syllabus. It is included for interest.

We might wonder, why go to all the trouble of defining the Borel σ-field? In other words, why can't we measure (the 'size' of) every possible subset of $\mathbb{R}$? We will answer these questions by constructing a strange looking set $\mathscr{V} \subseteq \mathbb{R}$; we will then show that it is not possible to define the Lebesgue measure of $\mathscr{V}$.

As usual, let $\mathbb{Q}$ denote the rational numbers. For any $x \in \mathbb{R}$ we define

$$\mathbb{Q}_x = \{x + q\,;\, q \in \mathbb{Q}\}. \tag{1.8}$$

Note that different x values may give the same $\mathbb{Q}_x$. For example, an exercise for you is to prove that $\mathbb{Q}_{\sqrt{2}} = \mathbb{Q}_{1+\sqrt{2}}$. You can think of $\mathbb{Q}_x$ as the set $\mathbb{Q}$ translated by x.

It is easily seen that $\mathbb{Q}_x \cap [0,1]$ is non-empty; just pick some rational q that is slightly less than x and note that $x + (-q) \in \mathbb{Q}_x \cap [0,1]$. Now, for each set $\mathbb{Q}_x$, we pick precisely one element $r \in \mathbb{Q}_x \cap [0,1]$ (it does not matter which element we pick). We write this number r as $r(\mathbb{Q}_x)$. Define

$$\mathscr{V} = \{r(\mathbb{Q}_x)\,;\, x \in \mathbb{R}\},$$

which is a subset of $[0,1]$. For each $q \in \mathbb{Q}$ define

$$\mathscr{V}_q = \{q + m\,;\, m \in \mathscr{V}\}.$$

Clearly $\mathscr{V} = \mathscr{V}_0$, and $\mathscr{V}_q$ is precisely the set $\mathscr{V}$ translated by q. Now, let us record some facts about $\mathscr{V}_q$.

Lemma 1.6.1 *It holds that*

1. *If $q_1 \neq q_2$ then $\mathscr{V}_{q_1} \cap \mathscr{V}_{q_2} = \emptyset$.*
2. *$\mathbb{R} = \bigcup_{q\in\mathbb{Q}} \mathscr{V}_q$.*
3. *$[0,1] \subseteq \bigcup_{q\in\mathbb{Q}\cap[-1,1]} \mathscr{V}_q \subseteq [-1,2]$.*

Before we prove this lemma, let us use it to show that $\mathscr{V}$ cannot have a Lebesgue measure. We will do this by contradiction: assume that $\lambda(\mathscr{V})$ is defined.

Since $\mathscr{V}$ and $\mathscr{V}_q$ are translations of each other, they must have the same Lebesgue measure. We write $c = \lambda(\mathscr{V}) = \lambda(\mathscr{V}_q)$, which does not depend on q. Let us write $\mathbb{Q}\cap[-1,1] = \{q_1, q_2, \ldots, \}$, which we may do because $\mathbb{Q}$ is countable. By parts (1) and (3) of Lemma 1.6.1 and property (M2) we have

$$\lambda\left(\bigcup_{q\in\mathbb{Q}\cap[-1,1]} \mathscr{V}_q\right) = \sum_{i=1}^{\infty} \lambda(\mathscr{V}_{q_i}) = \sum_{i=1}^{\infty} c.$$

Using the monotonicity property of measures (see Section 1.7) and part (3) of Lemma 1.6.1 we thus have

$$1 \leq \sum_{i=1}^{\infty} c \leq 3.$$

However, there is no value of c which can satisfy this equation! So it is not possible to define of the Lebesgue measure of $\mathscr{V}$. Since we know that we can define the Lebesgue measure on all Borel sets, the set $\mathscr{V}$ is not a Borel set.

The set $\mathscr{V}$ is known as a *Vitali set*. In higher dimensions even stranger things can happen with non-measurable sets; you might like to investigate the *Banach-Tarski paradox*.

Proof: [Of Lemma 1.6.1.] We prove the three claims in turn.

(1) Let $q_1, q_2 \in \mathbb{Q}$ be unequal. Suppose that some $x \in \mathscr{V}_{q_1} \cap \mathscr{V}_{q_2}$ exists – and we now look for a contradiction. By definition of $\mathscr{V}_q$ we have

$$x = q_1 + r(\mathbb{Q}_{x_1}) = q_2 + r(\mathbb{Q}_{x_2}). \tag{1.9}$$

By definition of $\mathbb{Q}_x$ we may write $r(\mathbb{Q}_{x_1}) = x_1 + q'_1$ for some $q'_1 \in \mathbb{Q}$, and similarly for x_2, so we obtain $x = q_1 + x_1 + q'_1 = q_2 + x_2 + q'_2$ where $q, q' \in \mathbb{Q}$. Hence, setting $q = q_2 - q_1 + q'_2 - q'_1 \in \mathbb{Q}$, we have $x_1 + q = x_2$, which by (1.8) means that $\mathbb{Q}_{x_1} = \mathbb{Q}_{x_2}$. Thus $r(\mathbb{Q}_{x_1}) = r(\mathbb{Q}_{x_2})$, so going back to (1.9) we obtain that $q_1 = q_2$. But this contradicts our assumption that $q_1 \neq q_2$. Hence x does not exist and $\mathscr{V}_{q_1} \cap \mathscr{V}_{q_2} = \emptyset$.

(2) We will show $\supseteq$ and $\subseteq$. The first is easy: since $\mathscr{V}_q \subseteq \mathbb{R}$ it is immediate that $\mathbb{R} \supseteq \bigcup_{q\in\mathbb{Q}} \mathscr{V}_q$.

Now take some $x \in \mathbb{R}$. Since we may take $q = 0$ in (1.8) we have $x \in \mathbb{Q}_x$. By definition of $r(\mathbb{Q}_x)$ we have $r(\mathbb{Q}_x) = x + q'$ for some $q' \in \mathbb{Q}$. By definition of $\mathscr{V}$ we have $r(\mathbb{Q}_x) \in \mathscr{V}$ and since $x = r(\mathbb{Q}_x) - q'$ we have $x \in \mathscr{V}_{-q'}$. Hence $x \in \bigcup_{q\in\mathbb{Q}} \mathscr{V}_q$.

(3) Since $\mathscr{V} \subseteq [0,1]$, we have $\mathscr{V}_q \cap [0,1] = \emptyset$ whenever $q \notin [-1,1]$. Hence, from part (2) and set algebra we have

$$\mathbb{R} \cap [0,1] \;=\; \left(\bigcup_{q\in\mathbb{Q}} \mathscr{V}_q\right) \cap [0,1] \;=\; \bigcup_{q\in\mathbb{Q}} \mathscr{V}_q \cap [0,1] \;=\; \bigcup_{q\in\mathbb{Q}\cap[-1,1]} \mathscr{V}_q \cap [0,1] \;\subseteq\; \bigcup_{q\in\mathbb{Q}\cap[-1,1]} \mathscr{V}_q.$$

This proves the first $\subseteq$ of (3). For the second simply note that $\mathscr{V} \subseteq [0,1]$ so $\mathscr{V}_q \subseteq [-1,2]$ whenever $q \in [-1,1]$. ■

Remark 1.6.2 We used the axiom of choice to define the function $r(\cdot)$.

1.7 Measures and limits

In this section we return to the consideration of arbitrary measure spaces (S, Σ, m). Let (A_n) be a sequence of sets in Σ. We say that it is *increasing* if $A_n \subseteq A_{n+1}$ for all $n \in \mathbb{N}$, and *decreasing* if $A_{n+1} \subseteq A_n$. When (A_n) is increasing, it is easily seen that (A_n^c) is decreasing.

When (A_n) is increasing, a useful technique is the *disjoint union trick* whereby we can write $\bigcup_{n=1}^\infty A_n = \bigcup_{n=1}^\infty B_n$ where the B_ns are all mutually disjoint by defining $B_1 = A_1$ and for $n > 1, B_n = A_n - A_{n-1}$. e.g. $\mathbb{R} = \bigcup_{n=1}^\infty [-n, n]$ and here $B_1 = [-1, 1], B_2 = [-2, -1) \cup (1, 2]$ etc.

Lemma 1.7.1 *Let $A_n \in \Sigma$ for all n. It holds that:*

1. *If (A_n) is increasing and $A = \bigcup_{n=1}^\infty A_n$ then $m(A) = \lim_{n\to\infty} m(A_n)$.*
2. *If (A_n) is decreasing and $A = \bigcap_{n=1}^\infty A_n$, and $m(A_1) < \infty$, then $m(A) = \lim_{n\to\infty} m(A_n)$.*

PROOF: We will prove the first claim here. The second claim can be deduced from the first, which is for you to do in Problem 1.7. We use the disjoint union trick and (M2) to find that

$$m(A) = m\left(\bigcup_{n=1}^\infty B_n\right) = \sum_{n=1}^\infty m(B_n) = \lim_{N\to\infty} \sum_{n=1}^N m(B_n) = \lim_{N\to\infty} m\left(\bigcup_{n=1}^N B_n\right) = \lim_{N\to\infty} m(A_N).$$

Here we use that $A_N = B_1 \cup B_2 \cup \cdots \cup B_N$. ■

Lemma 1.7.2 (Union bound) *If (A_n) is an arbitrary sequence of sets with $A_n \in \Sigma$ for all $n \in \mathbb{N}$ then*

$$m\left(\bigcup_{n=1}^\infty A_n\right) \le \sum_{n=1}^\infty m(A_n).$$

PROOF: From Problem 1.4, we have $m(A_1 \cup A_2) + m(A_1 \cap A_2) = m(A_1) + m(A_2)$ from which we deduce that $m(A_1 \cup A_2) \le m(A_1) + m(A_2)$. By induction we then obtain for all $N \ge 2$,

$$m\left(\bigcup_{n=1}^N A_n\right) \le \sum_{n=1}^N m(A_n).$$

Now define $X_N = \bigcup_{n=1}^N A_n$. Then $X_N \subseteq X_{N+1}$ and so (X_N) is increasing to $\bigcup_{n=1}^\infty X_n = \bigcup_{n=1}^\infty A_n$. By Lemma 1.7.1 we have

$$m\left(\bigcup_{n=1}^\infty A_n\right) = m\left(\bigcup_{n=1}^\infty X_n\right) = \lim_{N\to\infty} m(X_N) = \lim_{N\to\infty} m\left(\bigcup_{n=1}^N A_n\right) \le \lim_{N\to\infty} \sum_{n=1}^N m(A_n) = \sum_{n=1}^\infty m(A_n).$$

■

1.8 Null sets

Sets of measure zero play an important role in measure theory, and in advanced probability. In fact, there is a special terminology for them, which we introduce in this section. In a general measure space (S, Σ, m) we say that a set $E \in \Sigma$ is a *null set* if $m(E) = 0$. If we need to specify which measure is involved we might say e.g. m-null or Lebesgue null. We've already seen some examples of Lebesgue null sets in Section 1.5.

Lemma 1.8.1 *For each $n \in \mathbb{N}$ let E_n be a null set. Then $\bigcup_{n=1}^{\infty} E_n$ is a null set.*

PROOF: By Lemma 1.7.2 we have $m(\bigcup_{n=1}^{\infty} E_n) \leq \sum_{n=1}^{\infty} m(E_n) = 0$. ■

If a set $E \in \Sigma$ is such that its complement is null (i.e. $m(S \setminus E) = 0$) then we say that E has *full measure*. You can prove an analogue of Lemma 1.8.1 for sets of full measure in Exercise **1.8**.

We say that a property holds for *almost all* $x \in S$ if the set of x for which the property holds has full measure. This is best understood by example: in Section 1.5 we deduce that the rational numbers $\mathbb{Q}$ were a Lebesgue null subset of $\mathbb{R}$. Therefore the set of irrational numbers has full measure. We can rephrase this statement as 'almost all $x \in \mathbb{R}$ are irrational numbers'. If we needed to be specific that we meant to use Lebesgue measure, we might say 'Lebesgue almost all $x \in \mathbb{R}$ are irrational'.

Remark 1.8.2 ($\star$) Those of you taking courses in topics such as topology, metric spaces and functional analysis may wonder what relationships exist between sets of full measure and dense sets, when using the Borel σ-field on some metric or topological space as in Remark 3.3.4. The surprising answer is that in general a dense set might not have full measure, and a set of full measure might not be dense. The perspectives of measure theory (i.e. a set of full measure) and topology (i.e. a dense set) turn out to be quite different.

1.9 Product measures

We calculate areas of rectangles by multiplying products of lengths of their sides. This suggests trying to formulate a theory of products of measures. Let (S_1, Σ_1, m_1) and (S_2, Σ_2, m_2) be two measure spaces. Form the Cartesian product $S_1 \times S_2$. We can similarly try to form a product of σ-fields

$$\Sigma_1 \times \Sigma_2 = \{A \times B; A \in \Sigma_1, B \in \Sigma_2\},$$

but it turns out that $\Sigma_1 \times \Sigma_2$ is not a σ-field in general e.g. take $\Sigma_1 = \Sigma_2 = \mathbb{R}$ and note that $((0,1) \times (0,1))^c$ is not a rectangle. Instead the object we want is $\Sigma_1 \otimes \Sigma_2$, which is defined to be the smallest σ-field containing all the sets in $\Sigma_1 \times \Sigma_2$.

Theorem 1.9.1 *There exists a measure $m_1 \times m_2$ on $(S_1 \times S_2, \Sigma_1 \otimes \Sigma_2)$ such that*

$$(m_1 \times m_2)(A \times B) = m_1(A)m_2(B) \tag{1.10}$$

for all $A \in \Sigma_1, B \in \Sigma_2$.

Definition 1.9.2 The measure $m_1 \times m_2$ is called the product measure of m_1 and m_2.

We won't include a proof of Theorem 1.9.1 within this course. For example, consider $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$. We equip it with the Borel σ-field, $\mathcal{B}(\mathbb{R}^2) = \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$. Then the product Lebesgue measure $\lambda_2 = \lambda \times \lambda$ has the property that

$$\lambda_2((a,b) \times (c,d)) = (b-a)(d-c).$$

Of course, $(b-a)(d-c)$ is the area of the rectangle $(a,b) \times (c,d)$. In fact, from a mathematical point of view the measure λ_2 is the *definition* of area. Similarly, $\lambda_3 = \lambda \times \lambda \times \lambda$ is how we define volume, in three dimensions.

Remark 1.9.3 After thinking about $\lambda \times \lambda \times \lambda$, we might ask if, given measures m_1, m_2, m_3, we have $(m_1 \times m_2) \times m_3 = m_1 \times (m_2 \times m_3)$. It is true, but we won't prove it. Consequently we write both these as simply $m_1 \times m_2 \times m_3$, without any ambiguity.

We can go beyond 3 dimensions. Given n-measure spaces $(S_1, \Sigma_1, m_1), (S_2, \Sigma_2, m_2), \ldots, (S_n, \Sigma_n, m_n)$, we can iterate the above procedure to define the product σ-field $\Sigma_1 \otimes \Sigma_2 \otimes \cdots \otimes \Sigma_n$ and the product measure $m_1 \times m_2 \times \cdots \times m_n$ so that for $A_i \in \Sigma_i, 1 \le i \le n$,

$$(m_1 \times m_2 \times \cdots \times m_n)(A_1 \times A_2 \times \cdots \times A_n) = m_1(A_1)m_2(A_2)\cdots m_n(A_n).$$

In particular n-dimensional Lebesgue measure on $\mathbb{R}^n$ may be defined in this way.

Of course there are many measures that one can construct on $(S_1 \times S_2, \Sigma_1 \times \Sigma_2)$ and not all of these will be product measures. In probability spaces, product measures are closely related to the notion of independence, as we will see later. If you write $m_1 = m_2 = \mathbb{P}$ in (1.10) you might be able to see why.

1.10 Exercises on Chapter 1

1.1 Let $S = \{1, 2, 3, 4\}$. Show that the set $\mathcal{A} = \{\emptyset, \{1, 2, 3\}, \{4\}, \{1, 2\}, \{1, 2, 4\}, \{3\}, S\}$ is not a σ-field on S.

1.2 Let Σ_1 and Σ_2 be σ-fields of subsets of a set S. Note that

$$\Sigma_1 \cap \Sigma_2 = \{A \subseteq S\,;\, A \in \Sigma_1 \text{ and } A \in \Sigma_2\},$$
$$\Sigma_1 \cup \Sigma_2 = \{A \subseteq S\,;\, A \in \Sigma_1 \text{ or } A \in \Sigma_2\}.$$

(a) Show that $\Sigma_1 \cap \Sigma_2$ is a σ-field.

(b) Why is $\Sigma_1 \cup \Sigma_2$ not in general a σ-field? Give an example to demonstrate this.

1.3 Let (S, Σ) be a measurable space and let $X \in \Sigma$. Show that $\Sigma_X = \{A \cap X\,;\, A \in \Sigma\}$ is a σ-field on X.

1.4 (a) Let (S, Σ, m) be a measure space. Show that for all $A, B \in \Sigma$,

(i) $m(A \cup B) + m(A \cap B) = m(A) + m(B)$,

(ii) $m(A \cup B) \le m(A) + m(B)$.

(b) Use (a)(ii) to prove that if $A_1, A_2, \ldots, A_n \in \Sigma$ then $m\left(\bigcup_{i=1}^n A_i\right) \le \sum_{i=1}^n m(A_i)$.

1.5 Let (S, Σ, m) be a measure space.

(a) Let $k > 0$. Show that km is also a measure on (S, Σ) where for all $A \in \Sigma$,

$$(km)(A) = km(A).$$

Hence show that if m is a finite measure and $m(S) > 0$, then $\mathbb{P}(A) = \frac{m(A)}{m(S)}$ defines a probability measure for $A \in \Sigma$.

(b) Let $B \in \Sigma$. Show that $m_B(A) = m(A \cap B)$ for $A \in \Sigma$ defines a measure on (S, Σ).

(c) Suppose that m is a finite measure and $m(B) > 0$. Deduce that $\mathbb{P}_B$ is a probability measure where

$$\mathbb{P}_B(A) = \frac{m_B(A)}{m(B)}.$$

How does this relate to the notion of conditional probability?

1.6 Show that $\mathcal{B}(\mathbb{R})$ contains all closed intervals $[a, b]$, where $-\infty < a < b < \infty$.

1.7 (a) Let m be a finite measure on the measurable space (S, Σ).

(i) Let $A \in \Sigma$. Show that $m(S \setminus A) = m(S) - m(A)$.

(ii) Let $(A_n)_{n \in \mathbb{N}}$ be a decreasing sequence of sets in Σ. Show that $m(A_n) \to m(\cap_j A_j)$ as $n \to \infty$.

This question proves part 2 of Lemma 1.7.1. We have already proved part 1 and you should use part 1 to prove part 2.

(b) Take $S = \mathbb{N}$, let $\Sigma = \mathcal{P}(\mathbb{N})$ and let $m = \#$ be counting measure. Give an example of a decreasing sequence of subsets $A_n \subseteq \Sigma$ for which $m(\cap_j A_j) \ne \lim_{n\to\infty} m(A_n)$.

1.8 Let (S, Σ, m) be a measure space and for each $n \in \mathbb{N}$ let E_n have full measure. Show that $\bigcap_{n=1}^\infty E_n$ has full measure.

1.9 Show that if S is a set containing n elements, then the power set $\mathcal{P}(S)$ contains 2^n elements.

Hint: How many subsets are there of size r, for a fixed $1 \leq r \leq n$? The binomial theorem may also be of some use.

Challenge Questions

1.10 Let S be a finite set and Σ be a σ-field on S. Consider the set

$$\Pi = \{A \in \Sigma\,;\ \text{if } B \in \Sigma \text{ and } B \subseteq A \text{ then either } B = A \text{ or } B = \emptyset\}. \qquad (\star)$$

(a) Show that Π is a finite set.

(b) Using (a), let us enumerate the elements of Π as $\Pi = \{\Pi_1, \Pi_2, \cdots, \Pi_k\}$, where each Π_i is distinct from the others.

(i) Show that $\Pi_i \cap \Pi_j = \emptyset$ for $i \neq j$. *Hint: Could $\Pi_i \cap \Pi_j$ be an element of Π?*

(ii) Show that $\cup_{i=1}^k \Pi_i = S$. *Hint: If $C = S \setminus \cup_{i=1}^k \Pi_i$ is non-empty, is $C \in \Pi$?*

(iii) Let $A \in \Sigma$. Show that

$$A = \bigcup_{i \in I} \Pi_i$$

where $I = \{i = 1, \ldots, k\,;\ A \cap \Pi_i \neq \emptyset\}$.

1.11 Prove that both of the following claims are false.

(a) The Cantor set C contains an open interval $(a, b) \subseteq C$, where $a < b$.

(b) If a Borel set has non-zero Lebesgue measure then it contains an open interval.

Chapter 2

Real Analysis

In this chapter we cover a small amount of content that would (mostly) fit more naturally within a course on real analysis, but which was not covered as part of earlier courses due to lack of time. We'll first put together a natural way of dealing with limits that might take the values ±∞, and then we'll think about convergence of sequences of functions.

Given a set $A \subseteq \mathbb{R}$, or more generally a subset $A \subseteq S$ for some measurable space (S, Σ), we define the *indicator* function $\mathbb{1}_A : S \to \mathbb{R}$ by

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

We will study indicator functions a little in this chapter, in Exercise **2.5**. They will become very important to us in Chapters 3 and 4.

2.1 The extended reals

The *extended reals* are the real numbers, with ±∞ included,

$$\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}.$$

We have already seen that measures take values in $[0,\infty]$, and we'll need to learn how to deal naturally $\pm\infty$ within this course. Let's recap a few key definitions.

You should remember the definition of a convergent sequence $a_n \to a$ of real numbers:

$$\forall \epsilon > 0 \;\; \overbrace{\exists N \in \mathbb{N}\, \forall n \geq N,}^{\text{'eventually'}} \;\; |a_n - a| \leq \epsilon.$$

The best way to understand this is definition is by thinking about the terms marked as the word 'eventually'. By eventually we mean that, if you look far enough down the sequence, we will see some event happen for all remaining terms. So convergence means that *for all* $\epsilon > 0$*, eventually* $|a_n - a| \leq \epsilon$. Note that we can use $< \epsilon$ or $\leq \epsilon$ here – it is straightforward to show that these give equivalent definitions.

We also need to think about limits that have the value $\pm\infty$. In real analysis we might call this 'divergence' because $\pm\infty$ are not elements of $\mathbb{R}$, but when we work within the extended real numbers $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty, -\infty\}$ we can also use the term 'convergence to $\pm\infty$' for this case. We won't be fussy about that point of language within this course. The definition of $a_n \to \infty$ is that

$$\forall M \in (0,\infty)\; \exists N \in \mathbb{N}\; \forall n \geq N,\; a_n \geq M.$$

In words, *for all* $M \in (0,\infty)$*, eventually* $a_n \geq M$. For the case $a_n \to -\infty$ we use $M \in (-\infty, 0)$ and require instead that $a_n \leq M$.

You already know that a bounded monotone sequence of real numbers converges. Working in $\overline{\mathbb{R}}$ allows us to remove the boundedness requirement.

Lemma 2.1.1 *Let* (a_n) *be a monotone sequence of extended real numbers. Then there exists* $a \in \overline{\mathbb{R}}$ *such that* $a_n \to a$.

Proof: Without loss of generality we can assume that (a_n) is monotone increasing i.e. $a_{n+1} \geq a_n$ for all $n \in \mathbb{N}$, or else we could consider $(-a_n)$ in place of (a_n). Note that in real analysis you have already proved the case where (a_n) is a bounded sequence. If (a_n) is unbounded and increasing then $\sup_n a_n = \infty$, so for any $M \in (0, -\infty)$ there exists N with $a_N \geq M$, which implies that $M \leq a_N \leq a_{N+1} \leq a_{N+2} \leq \ldots$. In words, eventually $a_n \geq M$. Hence $a_n \to \infty$. ■

Remark 2.1.2 (⋆) The extended reals are a compact metric space, for example with the metric given in Exercise 2.8. This provides a better way to study convergence in $\overline{\mathbb{R}}$ where the points $\pm\infty$ do not need to be viewed as special cases. Metric spaces are not pre-requisite to our course, however for this reason we will tend to omit treating the 'special cases' $\pm\infty$ within proofs involving $\overline{\mathbb{R}}$, for example in Lemma 2.2.2.

We can do arithmetic in $\overline{\mathbb{R}}$ in natural ways, for example if $a \in \mathbb{R}$ then $a + \infty = \infty$. We have to be careful though, some objects like $\infty - \infty$ and $\frac{\infty}{\infty}$ do not make sense (formally, they are undefined) but otherwise it works as you'd expect. Such restrictions are necessary. They exist to prevent nonsensical calculations like $1 = \frac{\infty}{\infty} = \frac{\infty+\infty}{\infty} = \frac{\infty}{\infty} + \frac{\infty}{\infty} = 1 + 1 = 2$. You can find the precise rules for arithmetic in $\overline{\mathbb{R}}$ in Section 0.2.

2.1.1 The Borel σ-field and Lebesgue measure on $\overline{\mathbb{R}}$

Recall that we defined the Borel σ-field $\mathcal{B}(\mathbb{R})$ in Section 1.4. We extend the Borel σ-field to $\overline{\mathbb{R}}$ by defining

$$\mathcal{B}(\overline{\mathbb{R}})=\sigma(\mathcal{B}(\mathbb{R}),\{\infty\},\{-\infty\}),$$

that is the smallest σ-field that contains all elements of $\mathcal{B}(\mathbb{R})$ and the singleton sets $\{\infty\}$ and $\{-\infty\}$. The following lemma summarizes the connection.

Lemma 2.1.3 *Let $A\subseteq\overline{\mathbb{R}}$. Then $A\in\mathcal{B}(\overline{\mathbb{R}})$ if and only if $A\cap\mathbb{R}\in\mathcal{B}(\mathbb{R})$.*

Proof: We'll prove the forwards and backwards implications in turn. For the forwards implication, Note that the definition of $\mathcal{B}(\overline{\mathbb{R}})$ gives that $(a,b)\in\mathcal{B}(\overline{\mathbb{R}})$ for all $-\infty\leq a<b<\infty$. By Exercise **1.3** the set $\Sigma=\{A\cap\mathbb{R}\,;\,A\in\mathcal{B}(\overline{\mathbb{R}})\}$ is a σ-field on $\mathbb{R}$, and taking $A=(a,b)$ shows that Σ contains all open intervals of $\mathbb{R}$. Hence $\mathcal{B}(\mathbb{R})\subseteq\Sigma$, because $\mathcal{B}(\mathbb{R})$ is the smallest σ-field with that property. which is precisely what we want.

For the reverse implication, for any $A\subseteq\overline{\mathbb{R}}$ we can write $A=(A\cap\mathbb{R})\cup(A\cap\{\infty\})\cup(A\cap\{-\infty\})$. Note that $A\cap\{\infty\}$ is either empty or equal to $\{\infty\}$, which in either case is an element of $\mathcal{B}(\overline{\mathbb{R}})$. Similarly, $A\cap\{-\infty\}\in\mathcal{B}(\mathbb{R})$. Hence, if $A\cap\mathbb{R}\in\mathcal{B}(\mathbb{R})$ then $A\in\mathcal{B}(\overline{\mathbb{R}})$. ■

Recall that we defined Lebesgue measure $\lambda(A)$ for $A\in\mathcal{B}(\mathbb{R})$ in Section 1.5. We extend Lebesgue measure to $A\in\mathcal{B}(\overline{\mathbb{R}})$ by setting

$$\lambda(A)=\lambda(A\cap\mathbb{R}). \tag{2.1}$$

Lemma 2.1.3 ensures that $A\cap\mathbb{R}\in\mathcal{B}(\mathbb{R})$, so the right hand side of (2.1) may be used to define the left. In words, we don't put any weight at ±∞. It is straightforward to check that this gives a measure on $(\overline{\mathbb{R}},\mathcal{B}(\mathbb{R}))$, in very similar style to part (b) of Exercise **1.5**.

2.2 Liminf and limsup

The main difficulty with limits is that, in general, limits do not exist. Most sequences do not converge to anything. However there are two closely related concepts that *always* exist. They are much easier to work with but they take more care to define.

Let (a_n) be a sequence of real numbers. Note that the sequence $b_n = \sup_{k\geq n} a_k$ is monotone decreasing, because as n gets larger the set $\{a_k\,;\,k\geq n\}$ contains less terms. Lemma 2.1.1 implies that this sequence has a limit, with the caveat that the limit might be infinite. The sequence (b_n) is monotone decreasing, so its limit is equal to $\inf_{n\in\mathbb{N}} b_n$. With this is mind we make the following definition:

$$\limsup_{n\to\infty} a_n = \lim_{n\to\infty}\left(\sup_{k\geq n} a_k\right) = \inf_{n\in\mathbb{N}}\left(\sup_{k\geq n} a_k\right). \tag{2.2}$$

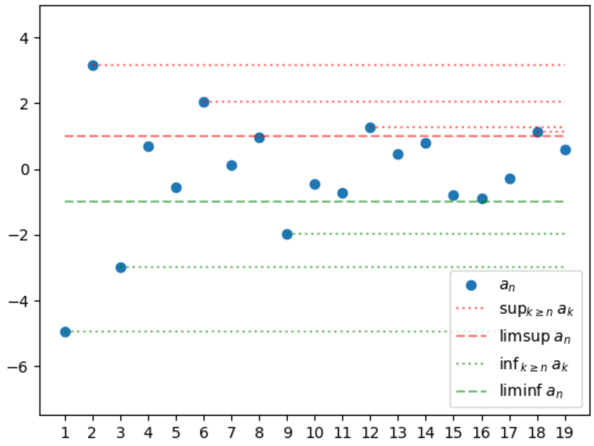
Heuristically, $\limsup_n a_n$ is the smallest value that the tail of the sequence (a_n) stays below.

We can do the same construction the other way up, which gives

$$\liminf_{n\to\infty} a_n = \lim_{n\to\infty}\left(\inf_{k\geq n} a_k\right) = \sup_{n\in\mathbb{N}}\left(\inf_{k\geq n} a_k\right). \tag{2.3}$$

Heuristically, $\liminf_n a_n$ is the largest value that the tail of the sequence (a_n) stays above. Note that (2.2) and (2.3) are always well defined, as extended real numbers, and that $\liminf_n a_n \leq \limsup_n a_n$.

Example 2.2.1 It is helpful to see a picture:



The sequence displayed is a sample of $a_n = \cos(n) + 10\frac{(-1)^n}{n}U_n$, where (U_n) are i.i.d. uniform random variables on $[0,1]$. This is chosen to make a clear picture. As $n\to\infty$ the $\cos(n)$ term will oscillate within $[-1,1]$ and the second term will tend to zero. Note that the dotted lines converge downwards to $\limsup_n a_n = 1$ (in red) and upwards to $\liminf_n a_n = -1$ (in green).

We will spend the rest of this section making connections between lim inf, lim sup and lim.

Lemma 2.2.2 *Let (a_n) be a sequence of extended reals. Then:*

1. *The sequence (a_n) converges if and only if $\liminf\limits_{n\to\infty} a_n = \limsup\limits_{n\to\infty} a_n$.*
2. *If (a_n) converges then $\liminf\limits_{n\to\infty} a_n = \limsup\limits_{n\to\infty} a_n = \lim\limits_{n\to\infty} a$.*

PROOF: Suppose first that (a_n) converges, say $a_n \to a$. We will consider the case $a \in \mathbb{R}$ here; the case $a = \pm\infty$ is similar and we will omit it, as discussed in Remark 2.1.2. For all $\epsilon > 0$ there exists $n \in \mathbb{N}$ such that $|a_k - a| \leq \epsilon$ for all $k \geq n$. Hence $a - \epsilon \leq a_k \leq a + \epsilon$ for all $k \geq n$, which implies that

$$a - \epsilon \;\leq\; \inf_{k \geq n} a_k \;\leq\; \sup_{k \geq n} a_k \;\leq\; a + \epsilon$$

Without loss of generality we may choose $n \geq \frac{1}{\epsilon}$. Letting $\epsilon \to 0$, upon which $n \to \infty$, gives that

$$a = \liminf_n a_n = \limsup_n a_n.$$

We have therefore proved both part 2 and the forwards implication of part 1.

We need to prove the reserve implication from part 1. Suppose that $\liminf_n a_n = \limsup_n a_n$ (and we don't yet know that (a_n) converges). For all $n \in \mathbb{N}$ we have

$$0 \leq a_n - \inf_{k \geq n} a_k \leq \sup_{k \geq n} a_k - \inf_{k \geq n} a_k. \tag{2.4}$$

Note also that

$$\lim_{n\to\infty} \left(\sup_{k \geq n} a_k - \inf_{k \geq n} a_k \right) = \limsup_{n\to\infty} a_n - \liminf_{n\to\infty} a_n = 0. \tag{2.5}$$

Combining (2.4) with (2.5) and using the sandwich rule, we have

$$\lim_{n\to\infty} \left(a_n - \inf_{k \geq n} a_k \right) = 0. \tag{2.6}$$

We can write $a_n = (a_n - \inf_{k \geq n} a_k) + \inf_{k \geq n} a_k$, and we know that both of these terms have a limit as $n \to \infty$. The first tends to zero by (2.6) and the second converges to $\liminf_n a_n$. From the algebra of limits we thus obtain that (a_n) converges and $\lim_n a_n = \liminf_n a_n$. Since $\liminf_n a_n = \limsup_n a_n$ was our assumption, this completes the proof. ■

Lemma 2.2.3 *Let (a_n) be a sequence of extended reals. Then $\limsup_n(-a_n) = -\liminf_n a_n$.*

PROOF: You already know from real analysis that $\sup_n(-a_n) = -\inf_n a_n$ for real a_n. This equation also holds for extended reals, but we'll omit checking the extra cases involving infinities here. The result follows from this along with (2.2) for lim sup, and with (2.3) for lim inf. ■

2.3 Convergence of functions

We've thought about convergence of (extended) real numbers in Sections 2.1 and 2.2. In this section we will think about convergence of functions. First we need to introduce *pointwise definitions* of functions. This is best done by example. If f and g are functions defined from $S \to \mathbb{R}$ then we define the function $f+g : S \to \mathbb{R}$ by setting $(f+g)(x) = f(x) + g(x)$. We can apply the same idea to fg to define the function $(fg)(x) = f(x)g(x)$, and so on.

Definition 2.3.1 Let f_n and f be functions defined from : $S \to \mathbb{R}$. We say that $f_n \to f$ *pointwise* if $f_n(x) \to f(x)$ as $n \to \infty$, for all $x \in S$.

Pointwise convergence is the simplest type of convergence of functions. You've already seen one other type: if $\sup_{x\in A} |f_n(x) - f(x)| \to 0$ as $n \to \infty$, then we say that $f_n \to f$ uniformly on the set A. We will use pointwise and uniform convergence within this course, but for us the most interesting type of convergence is something slightly different. Recall the term 'almost all' from Section 1.8. A property holds for almost all $x \in S$ if the set of x on which it fails is a null set.

Definition 2.3.2 Let (S, Σ, m) be a measure space and let f_n and f be functions defined from : $S \to \mathbb{R}$. We say that $f_n \to f$ *almost everywhere* if $f_n(x) \to f(x)$ for almost all $x \in S$.

We will sometimes abbreviate $f_n \to f$ almost everywhere as $f_n \overset{\text{a.e.}}{\to} f$. Unpacking the terminology in Definition 2.3.2, we have that $f_n \overset{\text{a.e.}}{\to} f$ if and only $m(\{x \in S\,;\, f_n(x) \nrightarrow f(x)\}) = 0$.

Convergence almost everywhere is very similar to pointwise convergence. The difference is that we allow $f_n(x) \to f(x)$ to fail on some null set of $x \in S$. This is much more natural from the perspective of measure theory, because we want to forget about things that have measure zero.

Example 2.3.3 Let $f_n : \mathbb{R} \to \mathbb{R}$ by $f_n(x) = e^{-nx^2}$ and let $f(x) = 0$. We take our measure space to be $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$, and note that as $n \to \infty$ we have $f_n(x) \to 0$ for all $x \in \mathbb{R}$ except $x = 0$ (at which $f_n(0) = 1$). The set $\{0\}$ is Lebesgue null, so $f_n \overset{\text{a.e.}}{\to} f$.

Lemma 2.3.4 *Let (S, Σ, m) be a measure space and let f_n and f be functions from $S \to \mathbb{R}$.*

1. *If $f_n \to f$ uniformly then $f_n \to f$ pointwise.*
2. *If $f_n \to f$ pointwise then $f_n \to f$ almost everywhere.*

PROOF: For the first claim, if $\sup_{x\in A} |f_n(x) - f(x)| \to 0$ then, for any $x \in A$, we have $f_n(x) \to f(x)$. For the second claim, pointwise convergence implies that the set $\{x \in S\,;\, f_n(x) \nrightarrow f(x)\}$ is empty, hence it has measure zero. ■

Example 2.3.3 shows that we can have convergence almost everywhere without having pointwise convergence. Exercise 2.7 gives an example of functions that converge pointwise but not uniformly.

2.4 Exercises on Chapter 2

2.1 Let (a_n) be a sequence of extended reals. Show that (a_n) is bounded if and only if we have $-\infty < \liminf_n a_n \leq \limsup_n a_n < \infty$.

Hint: Recall from real analysis that sequences which converge in $\mathbb{R}$ are necessarily bounded.

2.2 Let $f_n : \mathbb{R} \to \mathbb{R}$ by $f_n(x) = n\mathbb{1}_{[0,\frac{1}{n}]}(x)$. Show that $f_n \to 0$ almost everywhere.

2.3 Let (S, Σ, m) be a measure space. Let f, g and, for each $n \in \mathbb{N}$, f_n and g_n be functions from S to $\mathbb{R}$. Suppose that $f_n \to f$ almost everywhere, and $g_n \to g$ almost everywhere. Show that $f_n + g_n \to f + g$ almost everywhere, and $f_n g_n \to fg$ almost everywhere.

2.4 (a) For each $n \in \mathbb{N}$ let $a_n, b_n \in [0, \infty]$. Show that

$$\sum_{n=1}^{\infty} a_n + \sum_{n=1}^{\infty} b_n = \sum_{n=1}^{\infty} (a_n + b_n).$$

If $a_n, b_n \in \mathbb{R}$ and the summations converge in $\mathbb{R}$, implying absolute convergence because all terms are non-negative, then you already know this. The point of this question is to work in $\overline{\mathbb{R}}$.

(b) Let m and n be measures on (S, Σ). Deduce that $m + n$ is a measure on (S, Σ), where $(m + n)(A) = m(A) + n(A)$ for all $A \in \Sigma$.

(c) Let $S = \{x_1, x_2, \ldots, x_n\}$ be a finite set and $c_1, c_2, \ldots, c_n$ be non-negative numbers. Let

$$m = \sum_{i=1}^{n} c_i \delta_{x_i}.$$

(i) Show that m is a measure on $(S, \mathcal{P}(S))$.

(ii) What condition should be imposed on $\{c_1, c_2, \ldots, c_n\}$ for m to be a probability measure?

2.5 Let S be a set.

(a) Let $A, B \subseteq S$. Show that:

(i) $\mathbb{1}_{A \cup B} = \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_{A \cap B}$.

(ii) $\mathbb{1}_{A \cap B} = \mathbb{1}_A \mathbb{1}_B$.

(iii) If $B \subseteq A$ then $\mathbb{1}_{A \setminus B} = \mathbb{1}_A - \mathbb{1}_B$.

(b) Let (A_n) be a sequence of disjoint subset of S and set $A = \bigcup_{n=1}^{\infty} A_n$. Explain how the function $\sum_{n=1}^{\infty} \mathbb{1}_{A_n}$ is defined and show that $\mathbb{1}_A = \sum_{n=1}^{\infty} \mathbb{1}_{A_n}$.

2.6 Let (a_n) and (b_n) be a sequences of extended reals.

(a) Show that:

(i) $\limsup_n a_n + \limsup_n b_n \leq \limsup_n (a_n + b_n)$

(ii) $(\limsup_n a_n)(\limsup_n b_n) \leq \limsup_n a_n b_n$

(iii) $c \limsup_n a_n = \limsup_n (ca_n)$ for $c \in [0, \infty)$.

(b) Derive similar relationships for $\liminf$.

2.7 Let $f_n : \mathbb{R} \to \mathbb{R}$ be given by $f_n(x) = e^{-nx^2}$ and let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = \mathbb{1}_{\{0\}}(x)$. Show that $f_n \to f$ pointwise but not uniformly.

Challenge questions

2.8 *This question involves metric spaces and is off-syllabus for that reason.* ($\star$)

(a) Show that $(\overline{\mathbb{R}}, d)$ is a metric space, where $d(x,y) = |\arctan(x) - \arctan(y)|$. Here we set $\arctan(-\infty) = -1$ and $\arctan(\infty) = 1$.

(b) Show that $(\overline{\mathbb{R}}, d)$ is compact.

(c) Let (a_n) be a sequence of extended real numbers. Show that the following are equivalent:

(i) $a_n \to a$

(ii) If (a_{r_n}) is a convergent subsequence of (a_n) then $a_{r_n} \to a$.

2.9 Let (a_n) be any sequence within $\overline{\mathbb{R}}$ and let

$$\mathscr{L} = \{a \in \overline{\mathbb{R}}\,;\ \text{there exists a subsequence } (a_{r_n}) \text{ of } (a_n) \text{ with } a_{r_n} \to a\}.$$

Show that $\liminf_n a_n = \inf \mathscr{L}$ and $\limsup_n a_n = \sup \mathscr{L}$.

Part (c) of Exercise ***2.8*** *will help. If you prefer to avoid using that, you should impose the extra condition that $a_n \in \mathbb{R}$ is a bounded sequence and use* ***2.1****.*

Chapter 3

Measurable Functions

In this chapter we restrict ourselves to studying a particular kind of function, known as a measurable function. For measure theory, this is an important step, because it allows us to exclude some very strangely behaved examples that would disrupt our theory.

More specifically, in Section 1.6 we saw the existence subsets of $\mathbb{R}$ that were not measurable, with respect to Lebesgue measure. These sets had no meaningful concept of 'length'. It is clear that integration is closely connected to length, for example integrating the indicator function

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

in the case $A = [a, b]$ using a naive 'area under the curve' approach gives $\int_{\mathbb{R}} \mathbb{1}_A(x)\,dx = b - a = \lambda(A)$. If A is non-measurable then integrating the function $\mathbb{1}_A$ would be dangerously close to trying to measure the length of A, which we know we cannot do. We conclude that, just as with sets, we need a similar concept of a measurable *function.*

3.1 Overview

In this section, we deal with functions $f : S \to \mathbb{R}$, where (S, Σ) is a measurable space. The key definition is the following.

Definition 3.1.1 We say that $f : S \to \mathbb{R}$ is *measurable* if $f^{-1}(A) \in \Sigma$ whenever $A \in \mathcal{B}(\mathbb{R})$.

We will sometimes wish to restrict to the case $S = \mathbb{R}$, in which case we would normally take $\Sigma = \mathcal{B}(\mathbb{R})$ as the Borel sets. In this case, measurable functions $f : \mathbb{R} \to \mathbb{R}$ are often known as *Borel measurable.*

We can't yet explain exactly why Definition 3.1.1 is a sensible class of functions to be interested in. This will become clear in Section 4.1 when we begin to construct the Lebesgue integral. Note that the σ-field Σ is the key object in Definition 3.1.1. If we were to use a different σ-field Σ on the same set S, then we might change whether $f : S \to \mathbb{R}$ was measurable.

Example 3.1.2 Let $f : S \to \mathbb{R}$ and let $X \in \Sigma$. Then $\mathbb{1}_X$ is measurable. To see this let us write $f = \mathbb{1}_X$ and note that $f(x)$ takes the values 0 if $x \in X$ and 1 if $x \notin X$. Therefore we have

$$f^{-1}(A) = \begin{cases} \emptyset & \text{if } 0 \notin A \text{ and } 1 \notin A \\ S & \text{if } 0 \in A \text{ and } 1 \in A \\ A & \text{if } 0 \notin A \text{ and } 1 \in A \\ S \setminus A & \text{if } 0 \in A \text{ and } 1 \notin A. \end{cases}$$

In all cases we have that $f^{-1}(A) \in \Sigma$.

A interesting special case is provided the indicator function $\mathbb{1}_{\mathbb{Q}}$ of the rational numbers, defined from $\mathbb{R} \to \mathbb{R}$. This function is discontinuous at all points, because between any pair of rationals there is an irrational number, and vice versa. Since $\mathbb{Q} \in \mathcal{B}(\mathbb{R})$, the function $\mathbb{1}_{\mathbb{Q}}$ is Borel measurable.

Example 3.1.3 $(\star)$ It is possible to construct non-measurable functions, for example if we take $(S, \Sigma) = (\mathbb{R}, \mathcal{B}(\mathbb{R})$ and define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = \mathbb{1}_{\mathscr{V}}(x)$ where $\mathscr{V}$ is the non-measurable set constructed in Section 1.6. Then $f^{-1}(\{1\}) = \mathscr{V} \notin \mathcal{B}(\mathbb{R})$.

It is usually impossible to use Definition 3.1.1 to check directly that some function f is measurable, because $\mathcal{B}(\mathbb{R})$ is too big to check $f^{-1}(A)$ for all $A \in \mathcal{B}(\mathbb{R})$. In fact it is sufficient to check a much smaller class of subsets, for example using the following result.

Lemma 3.1.4 *Let $f : S \to \mathbb{R}$. Then the following statements are equivalent.*

1. *f is measurable.*
2. *$f^{-1}((a, \infty)) \in \Sigma$ for all $a \in \mathbb{R}$.*
3. *$f^{-1}([a, \infty)) \in \Sigma$ for all $a \in \mathbb{R}$.*
4. *$f^{-1}((-\infty, a)) \in \Sigma$ for all $a \in \mathbb{R}$.*
5. *$f^{-1}((-\infty, a]) \in \Sigma$ for all $a \in \mathbb{R}$.*

PROOF: It is immediate that part 1 ⇒ all of the other parts. We will show here that parts 2, 3, 4 and 5 are all equivalent to each other. Proof that these (all) imply part 1 can be found in Section 3.3, which is marked with a (Δ) for independent study.

Note first that part 2 ⇔ part 5, as $f^{-1}(A)^c = f^{-1}(A^c)$ and Σ is closed under taking complements. The fact that part 3 ⇔ part 4 is proved similarly. To see that part 2 ⇒ part 3 we use that $[a,\infty) = \bigcap_{n=1}^{\infty}(a-1/n,\infty)$ and so

$$f^{-1}([a,\infty)) = \bigcap_{n=1}^{\infty} f^{-1}((a-1/n,\infty))$$

and the result follows since Σ is closed under countable intersections. Similarly, to see that part 3 ⇒ part 2 we use that

$$f^{-1}((a,\infty)) = \bigcup_{n=1}^{\infty} f^{-1}([a+1/n,\infty))$$

and the fact that Σ is closed under countable unions. We thus have part part 5 ⇔ part 2 ⇔ part 3 ⇔ part 4. ■

There is nothing special about half-open intervals in Lemma 3.1.4. Lots of other types of subset of $\mathbb{R}$ will do and we will encounter a few more within this course, such as in Exercise **3.2**, or Lemma 3.3.5 within the independent reading.

Lemma 3.1.4 provides a direct way of showing that functions are measurable, but an indirect way is often easier: we can use measurable functions to construct more measurable functions. In fact, nearly everything that combines measurable functions together will create more measurable functions – much like the situation we already established for measurable sets. There are also similarities to the algebra of limits from real analysis, which gives the next theorem its name.

Recall the 'pointwise' notation for functions that we introduced in Section 2.3 e.g. $f+g$ means the function with values $(f+g)(x) = f(x)+g(x)$.

Theorem 3.1.5 (Algebra of measurable functions) *Let $f,g : S \to \mathbb{R}$ be measurable and let $\alpha \in \mathbb{R}$. The following functions are measurable:*

$$f+g, \qquad fg, \qquad \alpha f, \qquad 1/f, \qquad f \vee g, \qquad f \wedge g. \tag{3.1}$$

In the case of $1/f$ we must assume $f(x) \neq 0$ for all $x \in S$.

If $f_n : S \to \mathbb{R}$ for all $n \in \mathbb{N}$ then the following functions are measurable:

$$\inf_n f_n \qquad \sup_n f_n \qquad \liminf_{n\to\infty} f_n, \qquad \limsup_{n\to\infty} f_n, \qquad \lim_{n\to\infty} f_n \tag{3.2}$$

In the case of $\lim_n f_n$ we must assume that the limit exists (pointwise).

If $G : \mathbb{R} \to \mathbb{R}$ is Borel measurable then $G \circ f$, defined by $(G\circ f)(x) = G(f(x))$, is measurable.

Proof: The proof is in Section 3.4, which is marked with a (Δ) for independent study. ■

The functions in (3.1) are defined pointwise, as introduced in Section 2.3. The functions in (3.2) are also defined pointwise, for example $(\inf_n f_n)(x) = \inf_n f_n(x)$, and similarly for the others. Note that in the case of lim, the function $(\lim_n f_n)(x) = \lim_n f_n(x)$ is only defined if the limit exists for all x. Strictly, at this point we treat only real valued functions so we should require that the infs, sups, and so on are not $\pm\infty$, but we will remove this restriction in Section 3.6.

3.2 Borel measurable functions

To make use of the various parts of Theorem 3.1.5 we need some functions that we already know are measurable. Example 3.1.2 is a good start and you'll find some more examples within the exercises at the end of this chapter. In the special case of functions $f : \mathbb{R} \to \mathbb{R}$ the following lemma is very useful.

Lemma 3.2.1 *If $f : \mathbb{R} \to \mathbb{R}$ is continuous then f is Borel measurable.*

PROOF: The proof is in Section 3.3, which is marked with a (Δ) for independent study. ■

The most important fact to realize about functions $f : \mathbb{R} \to \mathbb{R}$ is that, although it is possible to construct non-measurable examples, essentially all examples of functions $f : \mathbb{R} \to \mathbb{R}$ that we encounter *in practice* are Borel measurable. We have now set up all these tools to prove this, for the functions that we commonly use. There are several examples in Exercise **3.1**. Here's one more.

Example 3.2.2 The function

$$f(x) = \begin{cases} \sin x & \text{if } x \in [\frac{-\pi}{2}, \frac{\pi}{2}], \\ 0 & \text{otherwise,} \end{cases}$$

is Borel measurable. To see this, note that the function sin is continuous on $\mathbb{R}$, and hence measurable by Lemma 3.2.1, which is a good start. To make the link to f we note that

$$f(x) = \mathbb{1}_{[\frac{-\pi}{2}, \frac{\pi}{2}]}(x)\sin(x).$$

Example 3.1.2 gives that $\mathbb{1}_{[-\frac{\pi}{2}, \frac{\pi}{2}]}$ is measurable, because intervals are Borel sets. Theorem 3.1.5 (in particular, the multiplication part) then gives that f is Borel measurable.

We could reach the same conclusion in many different ways. For example, the function

$$g(x) = \begin{cases} -1 & \text{if } x < \frac{-\pi}{2}, \\ \sin x & \text{if } x \in [\frac{-\pi}{2}, \frac{\pi}{2}], \\ 1 & \text{if } x > \frac{\pi}{2}, \end{cases}$$

is continuous and hence measurable by Lemma 3.2.1. We can write f as

$$f(x) = g(x) - \mathbb{1}_{(\frac{\pi}{2}, \infty)}(x) + \mathbb{1}_{(-\infty, \frac{-\pi}{2})}(x)$$

and Theorem 3.1.5 (this time, the addition part) tells us that f is Borel measurable.

3.3 Measurable functions and open sets (Δ)

In this section we prove the remaining part of Lemma 3.1.4, in particular that part (1) of that lemma is equivalent to the other parts. We will also prove Lemma 3.2.1. Note that this section is marked with a (Δ), meaning that it is off-syllabus for those taking MAS31002 and is independent reading for those taking MAS61022. Our arguments will use open subsets of $\mathbb{R}$. For purposes of this course we work from the following definition.

Definition 3.3.1 A set $O \subseteq \mathbb{R}$ is *open* if for every $x \in O$ there is an open interval $I \subseteq \mathbb{R}$ containing x, with $I \subseteq O$.

Some of you will have seen open sets in more general contexts e.g. metric or topological spaces. We won't use those more general contexts within this course, but if you are familiar with metric spaces you will know that some of the results in this section are true in greater generality than we include here.

It follows immediately from Definition 3.3.1 that every open interval in $\mathbb{R}$ is an open set. We might ask what other kinds of open subset we can find within $\mathbb{R}$. The following result gives a surprisingly clear answer, a consequence of which is that all open subsets of $\mathbb{R}$ are Borel sets.

Proposition 3.3.2 *Every open set O in $\mathbb{R}$ is a countable union of disjoint open intervals.*

PROOF: Note that a 'countable union' includes the case where we only need finitely many intervals. Let us first note that if O_i are opens sets for all $i \in I$ then (even if I is uncountable) the set $O = \cup_i O_i$ is open. See Exercise **3.7** for a proof of this fact.

For $x \in O$, let I_x be the union of all open intervals containing x for which $I_x \subseteq O$. Then I_x is open. Also, I_x is an interval, because if $a < b < c$ with $a, c \in I_x$ then there are open intervals $(a - \epsilon_1, x + \epsilon_2)$ and $(x - \epsilon_3, c + \epsilon_4)$ within I_x and b is within their union, so $b \in I_x$.

If $x, y \in O$ and $x \neq y$ then I_x and I_y are either disjoint or identical. To see this, note that if $I_x \cap I_y$ is non-empty then $I_x \cup I_y$ is a non-empty open interval contained within O, which implies $I_x \cup I_y$ is also contained within both I_x and I_y. Thus $I_x = I_y$.

However, there can only be countably many different I_x, because we can only fit at most countably many (non-empty) disjoint open intervals within $\mathbb{R}$. We now select a rational number $r(x)$ in every distinct I_x and rewrite O as the countable disjoint union over intervals I_x labelled by distinct rationals $r(x)$. ■

Lemma 3.3.3 *Let $\mathcal{O} = \sigma(O \subseteq \mathbb{R}\,;\, O \text{ is open})$ be the σ-field generated by the open subsets of $\mathbb{R}$. It holds that $\mathcal{O} = \mathcal{B}(\mathbb{R})$.*

PROOF: Recall that, by Definition 1.4.2, $\mathcal{B}(\mathbb{R})$ is the smallest σ-field that contains the open intervals (a, b) for $-\infty \leq a < b \leq \infty$. It follows immediately that $\mathcal{B}(\mathbb{R}) \subseteq \mathcal{O}$, because $\mathcal{O}$ is a σ-field containing all the such intervals and $\mathcal{B}(\mathbb{R})$ is the smallest such σ-field. To see the reverse inclusion, note by Proposition 3.3.2, every open set is an element of $\mathcal{B}(\mathbb{R})$. Thus $\mathcal{O} \subseteq \mathcal{B}(\mathbb{R})$, because $\mathcal{B}(\mathbb{R})$ therefore contains any σ-field that contains the open sets, and $\mathcal{O}$ is the smallest such σ-field. ■

Remark 3.3.4 $(\star)$ In advanced textbooks on measure theory, Lemma 3.3.3 is usually used as the definition of the Borel σ-field, because open sets make sense in a more general context than open intervals. In particular this definition makes sense for all metric spaces, and more generally for all topological spaces.

Lemma 3.3.5 *Let $f : S \to \mathbb{R}$. Then f is measurable if and only if $f^{-1}(O) \in \Sigma$ for all open sets O in $\mathbb{R}$.*

PROOF: We will prove the forwards and backwards implications in turn. Suppose first that f is measurable. Let $O \subseteq \mathbb{R}$ be open, and note that Proposition 3.3.2 implies that $O \in \mathcal{B}(\mathbb{R})$. Hence, by Definition 3.1.1 we have $f^{-1}(O) \in \Sigma$, as required.

For the reverse implication, suppose instead that $f^{-1}(O) \in \Sigma$ is for all open $O \subseteq \mathbb{R}$. Let $\mathcal{A} = \{E \subseteq \mathbb{R}\,;\, f^{-1}(E) \in \Sigma\}$. We will first show that $\mathcal{A}$ is a σ-field, by checking (S1)-(S3).

(S1): $\mathbb{R} \in \mathcal{A}$ as $S = f^{-1}(\mathbb{R})$.

(S2): If $E \in \mathcal{A}$ then $E^c \in \mathcal{A}$ since $f^{-1}(E^c) = f^{-1}(E)^c \in \Sigma$.

(S3): If (A_n) is a sequence of sets in $\mathcal{A}$ then $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$ since $f^{-1}\left(\bigcup_n A_n\right) = \bigcup_n f^{-1}(A_n) \in \Sigma$.

We are now ready to finish the proof. By our assumption, $O \in \mathcal{A}$ for all open $\mathcal{A} \subseteq \mathbb{R}$. Writing $\mathcal{O} = \sigma(O \subseteq \mathbb{R}\,;\, O \text{ is open})$ as in Lemma 3.3.3, by Lemma 1.2.6 we have that $\mathcal{O} \subseteq \mathcal{A}$, because $\mathcal{A}$ is a σ-field containing all the open subsets and $\mathcal{O}$ is the smallest such σ-field. By Lemma 3.3.3 we thus have $\mathcal{B}(\mathbb{R}) \subseteq \mathcal{A}$. By definition of $\mathcal{A}$, this gives that $f^{-1}(E) \in \Sigma$ for all $\mathcal{B}(\mathbb{R})$, so f is measurable. ■

PROOF OF LEMMA 3.1.4, PART 2 IMPLIES PART 1: With Lemma 3.3.5 we can finish the proof of Lemma 3.1.4. Using the notation from that lemma, assume that part 2 holds. From what we have already proved of Lemma 3.1.4, part 4 therefore also holds. By Proposition 3.3.2 we may write any open set O as $O = \cup_n (a_n, b_n)$ for some $-\infty \le a_n < b_n \le \infty$, where the union is countable. Hence,

$$f^{-1}(O) = \bigcup_n f^{-1}((a_n, b_n)) = \bigcup_n f^{-1}((-\infty, b_n)) \cap f^{-1}((a_n, \infty)).$$

The right hand side of the above is in Σ by parts 2 and 4, which means that $f^{-1}(O) \in \Sigma$ for any open set $O \subseteq \mathbb{R}$. From this, Lemma 3.3.5 gives that f is measurable. ■

The above completes the proof of Lemma 3.1.4, as promised from Section 3.1. We now move on to the proof of Lemma 3.2.1, starting with a proposition that links continuous functions to open sets.

Proposition 3.3.6 *A mapping $f : \mathbb{R} \to \mathbb{R}$ is continuous if and only if $f^{-1}(O)$ is open for every open set O in $\mathbb{R}$.*

PROOF: First suppose that f is continuous. Choose an open set O and let $a \in f^{-1}(O)$ so that $f(a) \in O$. Then there exists $\epsilon > 0$ so that $(f(a) - \epsilon, f(a) + \epsilon) \subseteq O$. By definition of continuity of f, for such an ϵ there exists $\delta > 0$ so that $x \in (a - \delta, a + \delta) \Rightarrow f(x) \in (f(a) - \epsilon, f(a) + \epsilon)$. But this tells us that $(a - \delta, a + \delta) \subseteq f^{-1}((f(a) - \epsilon, f(a) + \epsilon)) \subseteq f^{-1}(O)$. Since a is arbitrary we conclude that $f^{-1}(O)$ is open.

Conversely, suppose that $f^{-1}(O)$ is open for every open set O in $\mathbb{R}$. Choose $a \in \mathbb{R}$ and let $\epsilon > 0$. Then since $(f(a)-\epsilon, f(a)+\epsilon)$ is open so is $f^{-1}((f(a)-\epsilon, f(a)+\epsilon))$. Since $a \in f^{-1}((f(a)-\epsilon, f(a)+\epsilon))$

there exists $\delta > 0$ so that $(a-\delta, a+\delta) \subseteq f^{-1}((f(a)-\epsilon, f(a)+\epsilon))$. From here you can see that whenever $|x-a| < \delta$ we must have $|f(x)-f(a)| < \epsilon$. But then f is continuous at a and the result follows. ∎

Proof of Lemma 3.2.1: Let $f : \mathbb{R} \to \mathbb{R}$ be continuous and O be an arbitrary open set in $\mathbb{R}$. By Proposition 3.3.6 $f^{-1}(O)$ is an open set in $\mathbb{R}$. Hence, by Proposition 3.3.2 we have $f^{-1}(O) \in \mathcal{B}(\mathbb{R})$ for all open $O \subseteq \mathbb{R}$. Lemma 3.3.5 gives that f is measurable. ∎

3.4 Algebra of measurable functions (Δ)

In this section we prove Theorem 3.1.5. That is, we will show that, sums, products, limits etc of measurable functions are themselves measurable. The proof will be split across several lemmas. Throughout this section, (S, Σ) is a measurable space.

Lemma 3.4.1 *Let f, g be measurable functions from $S \to \mathbb{R}$. Then $f \vee g$ and $f \wedge g$ are measurable.*

Proof: Note that for all $c \in \mathbb{R}$ we have

$$(f \vee g)^{-1}((c, \infty)) = f^{-1}((c, \infty)) \cup g^{-1}((c, \infty))$$
$$(f \wedge g)^{-1}((c, \infty)) = f^{-1}((c, \infty)) \cap g^{-1}((c, \infty)).$$

We have that $f^{-1}((c, \infty))$ and $g^{-1}((c, \infty))$ are in Σ, hence the right hand side is in Σ. By parts 2 and 4 of Lemma 3.1.4, $f \vee g$ and $f \wedge g$ are measurable. ■

We write $\{f > g\} = \{x \in S\,;\, f(x) > g(x)\}$. We'll use similar notation for multiple inequalities of all types, and for constants. For example if $a, b \in \mathbb{R}$ then $\{a \leq f < c\} = \{x \in S\,;\, a \leq f(x) < c\}$.

Lemma 3.4.2 *Let f, g be measurable functions from $S \to \mathbb{R}$. Then:*

1. *$\{f > g\} \in \Sigma$.*
2. *$f + g$ and fg are measurable.*

Proof: We will use the results of Exercise 3.6 within this proof. For the first part, recall that the rational number $\mathbb{Q}$ are countable, and let $\{r_n, n \in \mathbb{N}\}$ be an enumeration of $\mathbb{Q}$. Recall also that there is a rational number between any two distinct real numbers. Hence,

$$\begin{aligned}\{f > g\} &= \bigcup_{n \in \mathbb{N}} \{f > r_n > g\} \\ &= \bigcup_{n \in \mathbb{N}} \{f > r_n\} \cap \{g < r_n\} \\ &= \bigcup_{n \in \mathbb{N}} f^{-1}((r_n, \infty)) \cap g^{-1}((-\infty, r_n)) \in \Sigma.\end{aligned}$$

The right hand side is in Σ, hence so is $\{f > g\}$.

For the second part, let us first consider $f + g$. From part (a) of Exercise 3.6, we have that $\alpha - g$ is measurable for all $\alpha \in \mathbb{R}$. Hence

$$(f + g)^{-1}((\alpha, \infty)) = \{f + g > \alpha\} = \{f > \alpha - g\},$$

is a measurable set by part 1 of the present lemma, so Lemma 3.4.2 gives that $f + g$ is measurable.

It remains to consider fg. Note that

$$fg = \frac{1}{4}[(f + g)^2 - (f - g)^2] \tag{3.3}$$

From part (c) of Exercise 3.6 we have that a composition of measurable functions is measurable. In particular, if $G(x) = x^2$ then $G \circ h = h^2$ is measurable whenever h is measurable. We have already shown that sums of measurable functions are measurable. Part (b) of Exercise 3.6 tells that αh is measurable whenever $\alpha \in \mathbb{R}$ and h is measurable. Thus $f + g$ is measurable, and $f - g = f + (-1) \times g$ is measurable, hence so are both of these squared, hence so is the right hand side of (3.3). This completes the proof. ■

Remark 3.4.3 ($\star$) Equation (3.3) is known as the *polarization identity* and is often useful when connecting multiplication and addition. A more general version of the same identity is used in functional analysis to connect norms with inner products.

Lemma 3.4.4 *For each $n \in \mathbb{N}$ let $f_n : S \to \mathbb{R}$ be a measurable function. Then, providing that all of the following are real valued:*

1. *$\inf_n f_n$ and $\sup_n f_n$ are measurable.*
2. *$\liminf_n f_n$ and $\limsup_n f_n$ are measurable.*
3. *If (f_n) converges pointwise to f as $n \to \infty$, then f is measurable.*

Proof: For part 1, note that for all $c \in \mathbb{R}$,

$$\left(\inf_{n\in\mathbb{N}} f_n\right)^{-1}([c,\infty)) = \{\forall n,\ f_n \geq c\} = \bigcap_{n\in\mathbb{N}} \{f_n \geq c\} = \bigcap_{n\in\mathbb{N}} f_n^{-1}([c,\infty)),$$
$$\left(\sup_{n\in\mathbb{N}} f_n\right)^{-1}((c,\infty)) = \{\exists n,\ f_n > c\} = \bigcup_{n\in\mathbb{N}} \{f_n > c\} = \bigcup_{n\in\mathbb{N}} f_n^{-1}((c,\infty)).$$

In both cases the right hand is in Σ, hence so is the left hand side. By Lemma 3.1.4, both $\inf_n f_n$ and $\sup_n f_n$ are measurable.

For part 2, recall from Section 2.2 that $\liminf_n f_n = \sup_n \inf_{k\geq n} f_k$ and $\limsup_n f_n = \inf_n \sup_{k\geq n} f_k$. By several applications of part 1, we have that $\liminf_n f_n$ and $\limsup_n f_n$ are measurable.

For part 3, if $f_n \to f$ pointwise then by Lemma 2.2.2 we have $f(x) = \liminf_n f_n(x)$ for all $x \in S$, so f is measurable by part 2. ■

3.5 Simple functions

In this section we are interested in the following class of functions.

Definition 3.5.1 Let (S, Σ) be a measurable space. We say that a function $f : S \to \mathbb{R}$ is *simple* if it has the form

$$f = \sum_{i=1}^{n} c_i \mathbb{1}_{A_i} \tag{3.4}$$

where $c_i \in \mathbb{R}$ and $A_i \in \Sigma$, with $A_i \cap A_j = \emptyset$ whenever $i \neq j$.

In words, a simple function is a (finite) linear combination of indicator functions of non-overlapping measurable sets. It follows from Example 3.1.2 and Theorem 3.1.5 that every simple function is measurable. Exercise **3.5** shows that sums and scalar multiples of simple functions are themselves simple, so the set of all simple functions forms a vector space.

Our next theorem explains the purpose of simple functions. They are a natural class of functions with which to approximate measurable functions. This will be a key ingredient of Lebesgue integration. Loosely, we will (in Chapter 4) specify how to integrate simple functions, and then use an approximation scheme to extend the same idea to measurable functions.

Our pointwise notation for functions, from Section 2.3, is also useful for inequalities involving functions. For example, we say that $f : S \to \mathbb{R}$ is *non-negative* if $f \geq 0$, meaning that $f(x) \geq 0$ for all $x \in S$. Similarly, we write $f \leq g$ to mean that $f(x) \leq g(x)$ for all x. It is easy to check that a simple function of the form (3.4) is non-negative if and only if $c_i \geq 0$ for all i.

Theorem 3.5.2 *Let $f : S \to \mathbb{R}$ be measurable and non-negative. Then there exists a sequence (s_n) of non-negative simple functions on S with $s_n \leq s_{n+1} \leq f$ for all $n \in \mathbb{N}$ so that (s_n) converges pointwise to f as $n \to \infty$. Moreover, if f is bounded then the convergence is uniform.*

PROOF: We split this into three parts.

Step 1: Construction of (s_n). Divide the interval $[0, n)$ into $n2^n$ subintervals $\{I_j, 1 \leq j \leq n2^n\}$, each of length $\frac{1}{2^n}$ by taking $I_j = \left[\frac{j-1}{2^n}, \frac{j}{2^n}\right)$. Let $E_j = f^{-1}(I_j)$ and $F_n = f^{-1}([n, \infty))$. Then $S = \bigcup_{j=1}^{n2^n} E_j \cup F_n$. We define for all $x \in S$

$$s_n(x) = \sum_{j=1}^{n2^n} \left(\frac{j-1}{2^n}\right) \mathbb{1}_{E_j}(x) + n\mathbb{1}_{F_n}(x).$$

Step 2: Properties of (s_n). For $x \in E_j, s_n(x) = \frac{j-1}{2^n}$ and $\frac{j-1}{2^n} \leq f(x) < \frac{j}{2^n}$ and so $s_n(x) \leq f(x)$. For $x \in F_n, s_n(x) = n$ and $f(x) \geq n$. So we conclude that $s_n \leq f$ for all $n \in \mathbb{N}$.

To show that $s_n \leq s_{n+1}$, fix an arbitrary j and consider $I_j = \left[\frac{j-1}{2^n}, \frac{j}{2^n}\right)$. For convenience, we write I_j as I and we observe that $I = I_1 \cup I_2$ where $I_1 = \left[\frac{2j-2}{2^{n+1}}, \frac{2j-1}{2^{n+1}}\right)$ and $I_2 = \left[\frac{2j-1}{2^{n+1}}, \frac{2j}{2^{n+1}}\right)$. Let $E = f^{-1}(I), E_1 = f^{-1}(I_1)$ and $E_2 = f^{-1}(I_2)$. Then $s_n(x) = \frac{j-1}{2^n}$ for all $x \in E, s_{n+1}(x) = \frac{j-1}{2^n}$ for all $x \in E_1$, and $s_{n+1}(x) = \frac{2j-1}{2^{n+1}}$ for all $x \in E_2$. It follows that $s_n \leq s_{n+1}$ for all $x \in E$. A similar (easier) argument can be used on F_n.

Step 3: Convergence of (s_n). Fix any $x \in S$. Since $f(x) \in \mathbb{R}$ there exists $n_0 \in \mathbb{N}$ so that $f(x) \leq n_0$. Then for each $n > n_0, f(x) \in I_j$ for some $1 \leq j \leq n2^n$, i.e. $\frac{j-1}{2^n} \leq f(x) < \frac{j}{2^n}$. But $s_n(x) = \frac{j-1}{2^n}$ and so $|f(x) - s_n(x)| < \frac{1}{2^n}$ and pointwise convergence follows. Note that if f is bounded we can find $n_0 \in \mathbb{N}$ so that $f(x) \leq n_0$ for all $x \in \mathbb{R}$. Then the argument just given yields $|f(x) - s_n(x)| < \frac{1}{2^n}$ for all $x \in \mathbb{R}$ which gives uniform convergence ∎

3.6 Extended real functions

Let (S, Σ, m) be a measure space. An *extended real function* on S is a mapping $f : S \to \overline{\mathbb{R}}$, which might take the values $\pm\infty$. We introduced extensions of the Borel σ-field and Lebesgue measure to $\overline{\mathbb{R}}$ in Section 2.1.1. Using these extensions, all of theory that we have developed in this Chapter can also be made to work for extended real functions. In that context, in Lemma 3.1.4 we would use half-open intervals that contained $\pm\infty$ e.g. $(a, \infty]$ instead of (a, ∞), but everything else works essentially the same.

Remark 3.6.1 $(\star)$ To extend Section 3.4 to extended real functions we would need to use continuity and open sets involving $\pm\infty$. This happens naturally when treating $\overline{\mathbb{R}}$ as a metric space (as in Exercise **2.8**), but this course does not assume knowledge of metric spaces and such arguments are outside of what we can cover here.

3.7 Exercises on Chapter 3

3.1 Show that the following functions are Borel measurable, as maps from $\mathbb{R}$ to itself.

(a) The constant function $f(x) = \alpha$, where $\alpha \in \mathbb{R}$.

(b) $g(x) = \begin{cases} 0 & \text{for } x < 0 \\ e^x & \text{for } x \geq 0. \end{cases}$

(c) $h(x) = \sin(\cos(x))$

(d) $i(x) = \sin(x^2 \mathbb{1}_{[0,\infty)}(x))$

3.2 Let (S, Σ) be a measurable space and let $f : S \to \mathbb{R}$. Show that f is measurable if and only if $f^{-1}((a, b)) \in \Sigma$ for all $-\infty \leq a < b \leq \infty$.

3.3 Let (S, Σ) be a measurable space and let $f : S \to \mathbb{R}$ be measurable. Show that $|f|$ is measurable.

3.4 Let $f : \mathbb{R} \to \mathbb{R}$ and let $\alpha \in \mathbb{R}$.

(a) Suppose that f is Borel measurable. Show that the mapping $h : \mathbb{R} \to \mathbb{R}$ is measurable, where $h(x) = f(x + \alpha)$.

(b) Suppose that f is differentiable. Explain why both f and its derivative f' are measurable functions.

(c) Suppose that f is monotone increasing. Show that f is measurable.
Hint: Show that $f^{-1}((c, \infty))$ is an interval. Recall that $I \subseteq \mathbb{R}$ is an (open, closed or half-open) interval of $\mathbb{R}$ if, whenever $a, b \in I$ and $a < c < b$ we have $c \in I$.

3.5 Let (S, Σ) be a measurable space. Show that the set V of simple functions $f : S \to \mathbb{R}$, with pointwise addition and scalar multiplication, is a real vector space.

3.6 (Δ) Let (S, Σ) be a measurable space and $f : S \to \mathbb{R}$ be measurable. Let $\alpha \in \mathbb{R}$.

In this question you may not use the algebra of measurable functions (Theorem 3.1.5) or any of the results in Section 3.4. We use the results (a)-(c) in the proof of Theorem 3.1.5.

(a) Show that $g = f + \alpha$ is measurable.

(b) Show that $g = \alpha f$ is measurable.

(c) Let $G : \mathbb{R} \to \mathbb{R}$ be Borel measurable. Show that $G \circ f : S \to \mathbb{R}$ is measurable, where $(G \circ f)(x) = G(f(x))$.

3.7 (Δ) Recall the definition of an open set, from Definition 3.3.1.

(a) Let O_1 and O_2 be open subsets of $\mathbb{R}$. Show that $O_1 \cup O_2$ and $O_1 \cap O_2$ are also open.

(b) For each $n \in \mathbb{N}$ let O_n be an open subset of $\mathbb{R}$. Consider the following claims:

(i) $A = \bigcup_{n\in\mathbb{N}} O_n$ is open.

(ii) $B = \bigcap_{n\in\mathbb{N}} O_n$ is open.

Which of these claims are true? Give a proof or a counterexample in each case.

(c) A set $C \subseteq \mathbb{R}$ is said to be *closed* if $\mathbb{R} \setminus C$ is open. Which of your results from parts (a) and (b) hold for closed sets?

Challenge questions

3.8 A function $f : \mathbb{R} \to \mathbb{R}$ is said to be *upper-semicontinuous* at $x \in \mathbb{R}$, if given any $\epsilon > 0$ there exists $\delta > 0$ so that $f(y) < f(x) + \epsilon$ whenever $|x - y| < \delta$.

(a) Show that $f = \mathbb{1}_{[a,\infty)}$ (where $a \in \mathbb{R}$) is upper-semicontinuous for all $x \in \mathbb{R}$,

(b) Deduce that the floor function $f(x) = \lfloor x \rfloor$, which is equal to the greatest integer less than or equal to x, is upper-semicontinuous at all $x \in \mathbb{R}$.

(c) Show that if f is upper-semicontinuous for all $x \in \mathbb{R}$ then f is measurable.

Chapter 4

Lebesgue Integration

The concept of integration as a technique that both acts as an inverse to the operation of differentiation and also computes areas under curves goes back to the origin of the calculus and the work of Isaac Newton (1643-1727) and Gottfried Leibniz (1646-1716). It was Leibniz who introduced the $\int \cdots dx$ notation. The first rigorous attempt to understand integration as a limiting operation within the spirit of analysis was due to Bernhard Riemann (1826-1866). The approach to *Riemann integration* that is often taught (as in MAS2004/2009) was developed soon after by Jean-Gaston Darboux (1842-1917). At the time it was developed, this theory seemed to be all that was needed, but as the 19th century drew to a close, some problems appeared:

- One of the main tasks of integration is to recover a function f from its derivative f'. But some functions were discovered for which f' existed and was bounded, but where f' was not Riemann integrable.
- Suppose (f_n) is a sequence of functions converging pointwise to f. People wanted a useful set of conditions under which
$$\int f(x)\,dx = \lim_{n\to\infty}\int f_n(x)\,dx. \tag{4.1}$$
but weren't able to find any suitable conditions. Problem 4.18 illustrates some of the difficulties here; it gives an example of f_n, f such that $f_n(x) \to f(x)$ for all x, but in which (4.1) fails.
- Riemann integration was limited to computing integrals over $\mathbb{R}^n$ with respect to Lebesgue measure. Although it was not yet apparent, the emerging theory of probability would require the calculation of expectations of random variables X using the formula $\mathbb{E}(X) = \int_\Omega X(\omega)d\mathbb{P}(\omega)$. This requires a version of integration that works on a general measure space.

A new approach to integration was needed. In this chapter, we'll study *Lebesgue integration*, which allow us to investigate $\int_S f(x)\,dm(x)$ where $f : S \to \mathbb{R}$ is a 'suitable' measurable function defined on a general measure space (S, Σ, m). It was developed by Henri Lebesgue (pronounced 'Leb-eyg') and first published in 1902.

We will see that if we take m to be Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ then we recover the familiar integral $\int_{\mathbb{R}} f(x)\,dx$ but we will now be able to integrate a much bigger class of functions than Riemann and Darboux could. Most importantly, the Lebesgue integral solves all three of the issues discussed above.

4.1 The Lebesgue integral for simple functions

We'll present the construction of the Lebesgue integral in three steps. We'll work over a general measure space (S, Σ, m) for most of Chapter 4, and we'll integrate functions $f : S \to \mathbb{R}$. The first step will involve simple functions. In the second step we extend to non-negative measurable functions, and the final step we extend to what are known as 'Lebesgue integrable' functions.

Definition 4.1.1 (Lebesgue Integral, Step 1) If $A \in \Sigma$ then the integral of the indicator function $\mathbb{1}_A$ is defined as

$$\int_S \mathbb{1}_A \, dm = m(A). \tag{4.2}$$

More generally, if $f = \sum_{i=1}^n c_i \mathbb{1}_{A_i}$ is a simple function, then we define

$$\int_S f \, dm = \sum_{i=1}^n c_i m(A_i). \tag{4.3}$$

Note that $m(A_i)$ might be infinite, so $\int_S f dm \in [0, \infty]$.

Note that we can represent f in more than one way as a simple function. For example if $f = \mathbb{1}_{[0,1]}$ then also $f = \mathbb{1}_{[0,\frac{1}{2})} + \mathbb{1}_{[\frac{1}{2},1]}$. It is easy to guess that the value of (4.3) does not depend on the choice of representation. We'll omit a formal proof of this fact. Note also that equations (4.2) and (4.3) are consistent with each other, in the sense that if we take $n = 1$ and $c_1 = 1$ in (4.3) then we obtain (4.2).

When we work with the theory of integration we will tend to write $\int_S f \, dm$ for the Lebesgue integral of f. We will sometimes use the shorthand notation

$$\mathcal{I}(f) = \int_S f \, dm,$$

to make our proofs easier to read. For calculations it is often more helpful to write $\int_S f(x) \, dm(x)$, which is closer to the notation you've used before in the case $S = \mathbb{R}$. We'll come back to this point after we've reached Step 2 of the construction.

In each step of defining the Lebesgue integral, we'll establish some useful properties of the integral. Because we expand the amount of functions we can integrate at each step, this will mean that we carry several properties with us as we go, and we do some work in each step to upgrade them. We begin this process with the next lemma.

Lemma 4.1.2 *If f and g are simple functions then:*

1. *Linearity: for all $\alpha, \beta \in \mathbb{R}$*

$$\int_S (\alpha f + \beta g) \, dm = \alpha \int_S f \, dm + \beta \int_S g \, dm,$$

2. *Monotonicity:*

$$f \leq g \quad \Rightarrow \quad \int_S f \, dm \leq \int_S g \, dm.$$

Proof: Let us write $f = \sum_{i=1}^n c_i \mathbb{1}_{A_i}$. Note that we can assume without loss of generality that $\bigcup_{i=1}^n A_i = S$ by including an extra term with $c_{n+1} = 0$ and $A_{n+1} = S \setminus (\bigcup_{i=1}^n A_i)$ into the summation. Similarly, write $g = \sum_{j=1}^m d_j \mathbb{1}_{B_j}$ where $\bigcup_{i=1}^n B_i = S$. By the definition of simple functions we have $A_i \cap A_j = \emptyset$ and $B_i \cap B_j = \emptyset$ for all $i \neq j$.

We have

$$\begin{aligned}f=\sum_{i=1}^n c_i\mathbb{1}_{A_i\cap S}&=\sum_{i=1}^n c_i\mathbb{1}_{A_i\cap\bigcup_{j=1}^m B_j}\\&=\sum_{i=1}^n c_i\mathbb{1}_{A_i}\mathbb{1}_{\bigcup_{j=1}^m B_j}=\sum_{i=1}^n c_i\mathbb{1}_{A_i}\sum_{j=1}^m\mathbb{1}_{B_j}=\sum_{i=1}^n\sum_{j=1}^m c_i\mathbb{1}_{A_i}\mathbb{1}_{B_j}=\sum_{i=1}^n\sum_{j=1}^m c_i\mathbb{1}_{A_i\cap B_j}.\end{aligned}$$

Here, the first line uses $\bigcup_{i=1}^n A_i=S$ and the second line uses the results of Exercise **2.5**. We can obtain a similar expression for g, giving $g=\sum_{i=1}^n\sum_{j=1}^m+d_j\mathbb{1}_{A_i\cap B_j}$. It follows that

$$\alpha f+\beta g=\sum_{i=1}^n\sum_{j=1}^m(\alpha c_i+\beta d_j)\mathbb{1}_{A_i\cap B_j}. \tag{4.4}$$

Thus

$$\begin{aligned}\mathcal{I}(\alpha f+\beta g)&=\sum_{i=1}^n\sum_{j=1}^m(\alpha c_i+\beta d_j)m(A_i\cap B_j)\\&=\alpha\sum_{i=1}^n c_i\sum_{j=1}^m m(A_i\cap B_j)+\beta\sum_{j=1}^n d_i\sum_{i=1}^m m(A_i\cap B_j)\\&=\alpha\sum_{i=1}^n c_im\left(A_i\cap\bigcup_{j=1}^m B_j\right)+\beta\sum_{j=1}^m d_jm\left(\bigcup_{i=1}^n A_i\cap B_j\right)\\&=\alpha\sum_{i=1}^n c_im(A_i\cap S)+\beta\sum_{j=1}^m d_jm(B_j\cap S)\\&=\alpha\sum_{i=1}^n c_im(A_i)+\beta\sum_{j=1}^m d_jm(B_j)\\&=\alpha\mathcal{I}(f)+\beta\mathcal{I}(g).\end{aligned}$$

This proves linearity. For monotonicity, note that by linearity we have $\mathcal{I}(g)=\mathcal{I}(f)+\mathcal{I}(g-f)$. Putting $\alpha=1$ and $\beta=-1$ into (4.4), we have that we have that $g-f$ is a non-simple function. If $f\le g$ then $g-f\ge 0$, which means that we can write $g-f$ in the form $g-f=\sum_{i=1}^l e_i\mathbb{1}_{F_i}$ where $e_i\ge 0$, from which (4.3) gives that $\mathcal{I}(g-f)\ge 0$. Hence $\mathcal{I}(f)\le\mathcal{I}(g)$, as required. ■

4.1.1 Integration over subsets

Integrals over the real numbers are commonly written in the form $\int_a^b$, which denotes integration over the interval $[a,b]\subseteq\mathbb{R}$. We now introduce some notation for this, in the general case.

Definition 4.1.3 If $A\in\Sigma$, whenever $\int_S f dm$ is defined for some $f:S\to\mathbb{R}$ we define

$$\mathcal{I}_A(f)=\int_A f\,dm=\int_S \mathbb{1}_A f\,dm.$$

We call $\mathcal{I}_A(f)$ the integral of f over the set A. In general there is no guarantee that $I_A(f)$ is defined for some function f. We need f to be one of the types of functions that we work with in the steps used to define the integral.

The following lemma is another property of the Lebesgue integral that we will carry with us as we build up the definition.

Lemma 4.1.4 *Let $f:S\to\mathbb{R}$ be a simple function. Then $\nu:\Sigma\to\mathbb{R}$ by*

$$\nu(X)=\int_X f\,dm$$

is a measure.

PROOF: Let us write $f=\sum_{i=1}^n c_i\mathbb{1}_{A_i}$. Note that $\mathbb{1}_X f=\sum_{i=1}^n c_i\mathbb{1}_{A_i}\mathbb{1}_X=\sum_{i=1}^n c_i\mathbb{1}_{A_i\cap X}$, where we have used the identity $\mathbb{1}_A\mathbb{1}_B=\mathbb{1}_{A\cap B}$ from Exercise **2.5**. By Definition 4.1.3 and (4.3) we have

$$\int_X f\,dm=\sum_{i=1}^n c_i m(A_i\cap X). \tag{4.5}$$

By part (b) of Exercise **1.5** we have that $X\mapsto m(A_i\cap X)$ defines a measure, for all i. Using part (a) of the same exercise, $X\mapsto c_i m(A_i\cap X)$ also defines a measure. In Exercise **2.4** we showed that a finite sum of measures was also a measure, hence in fact the right hand side of (4.5) defines a measure. This completes the proof. ∎

4.2 The Lebesgue integral for non-negative measurable functions

We continue to work over a general measure space (S, Σ, m). When $f : S \to \mathbb{R}$ is measurable and non-negative, it is tempting to try and take advantage of Theorem 3.5.2 by defining "$\int_S f dm = \lim_{n\to\infty} \int_S s_n dm$", where (s_n) is increasing sequence of simple functions, converging pointwise to f. There is a problem with this idea: many different choices of simple functions could be used for s_n, which risks making the limiting integral depend on that choice. We'd need to show that didn't happen – which it doesn't, but proving that turns out to be very difficult. Lebesgue's key idea was to work around the problem, using the weaker notion of the supremum to 'approximate f from below' as follows.

We say that function $f : S \to \mathbb{R}$ is non-negative if $f \geq 0$. Equivalently, if $f : S \to [0, \infty)$. In this section we focus on non-negative measurable functions.

Definition 4.2.1 (Lebesgue Integral, Step 2) Let $f : S \to [0, \infty)$ be measurable. The integral of f is defined as

$$\int_S f\,dm = \sup\left\{\int_S s\,dm\,;\, s \text{ is a simple function and } 0 \leq s \leq f\right\}. \tag{4.6}$$

With this definition $\int_S f\,dm \in [0, \infty]$.

We would like to upgrade the monotonicity and linearity properties from Lemma 4.1.2. The use of the sup makes it easy to prove some properties but hard to prove others, so we'll only make partial progress for now. In particular, we'll have to postpone the upgrade of linearity until the next section.

Lemma 4.2.2 *Let $f, g : S \to [0, \infty)$ be measurable functions.*

1. *If $f \leq g$ then $\int_S f dm \leq \int_S g dm$.*
2. *If $A, B \in \Sigma$ with $A \subseteq B$ then $\int_A f\,dm \leq \int_B f\,dm$.*

PROOF: For part 1, if $f \leq g$ then any simple function s with $s \leq f$ also satisfies $s \leq g$. Hence

$$\begin{aligned}\int_S f\,dm &= \sup\left\{\int_S s\,dm\,;\, s \text{ is simple, } 0 \leq s \leq f\right\}\\ &\leq \sup\left\{\int_S s\,dm\,;\, s \text{ is simple, } 0 \leq s \leq g\right\}\\ &= \int_S g\,dm.\end{aligned}$$

Part 2 follows by noting that $\mathbb{1}_A \leq \mathbb{1}_B$, hence $\mathbb{1}_A f \leq \mathbb{1}_B f$ and applying part 1. ■

Lemma 4.2.3 (Markov's Inequality) *Let $f : S \to [0, \infty)$ be measurable and let $c > 0$. Then*

$$m(\{x \in S; f(x) \geq c\}) \leq \frac{1}{c}\int_S f\,dm$$

PROOF: Let $E = \{x \in S; f(x) \geq c\}$. Note that $E = f^{-1}([c, \infty)) \in \Sigma$ as f is measurable. By part 1 of Lemma 4.2.2 and the linearity property (for integrals of simple functions) from Lemma 4.1.2,

$$\int_S f\,dm \geq \int_S \mathbb{1}_E f\,dm \geq \int_S c\mathbb{1}_E\,dm = c\int_S \mathbb{1}_E\,dm = cm(E),$$

and the result follows. ■

You have already seen Markov's inequality in the context of probability theory, that is $\mathbb{P}[X \geq c] \leq \frac{1}{c}\mathbb{E}[X]$ for $c > 0$ and a random variable $X \geq 0$. In Exercise 5.1 we will see how the probabilistic version can be deduced from Lemma 4.2.3. For now, it is a useful tool within our development of integration, because it allows us to control how often some function f takes large values.

4.2.1 Almost everywhere equality

In Section 2.3 we introduce the concept of convergence almost everywhere. The idea behind $\stackrel{a.e.}{\to}$ was that we can allow pointwise convergence to fail on a null set, and from the point of measure theory that will tend not to matter. We will see this in practice within the next section, but first let us note that we can also apply 'almost everywhere' to the concept of equality, as follows.

Definition 4.2.4 Let $f, g : S \to \mathbb{R}$ be measurable. We say that $f = g$ *almost everywhere* if $f(x) = g(x)$ for almost all $x \in S$.

We will often abbreviate $f = g$ almost everywhere to $f \stackrel{a.e.}{=} g$. In symbols, $f \stackrel{a.e.}{=} g$ means that $m(\{x \in S; f(x) \neq g(x)\} = 0$. In Problem 4.11 you can show that $\stackrel{a.e.}{=}$ is an equivalence relation on the set of measurable functions from S to $\mathbb{R}$. You should think of $f \stackrel{a.e.}{=} g$ as saying 'as far as measure theory is concerned, f and g might as well be equal'.

Lemma 4.2.5 *Let $f : S \to [0, \infty)$ be measurable. If $\int_S f\,dm = 0$ then $f \stackrel{a.e.}{=} 0$.*

PROOF: Suppose that $\int_S f\,dm = 0$. Let $A = \{x \in S; f(x) \neq 0\}$ and for each $n \in \mathbb{N}$, $A_n = \{x \in S; f(x) \geq 1/n\}$. Note that $A = \bigcup_{n=1}^{\infty} A_n$. By Lemma 1.7.2 we have we have $m(A) \leq \sum_{n=1}^{\infty} m(A_n)$, hence it is sufficient to show that $m(A_n) = 0$ for all $n \in \mathbb{N}$. By Lemma 4.2.3 we have $m(A_n) \leq n\int_S f dm = 0$, which completes the proof. ■

Lemma 4.2.6 *Let $f, g : S \to [0, \infty)$ be measurable. If $f \stackrel{a.e.}{=} g$ then $\int_S f\,dm = \int_S g\,dm$.*

PROOF: Let us write $E = \{x \in S\,;\, f(x) \neq g(x)\}$ and note that $m(E) = 0$. Let $s = \sum_{i=1}^{n} c_i \mathbb{1}_{A_i}$ be a simple function such that $0 \leq s \leq f$. We claim that $s^* = \sum_{i=1}^{n} c_i \mathbb{1}_{A_i \setminus E}$ is a simple function: this follows because we have $A_i, E \in \Sigma$ so $A_i \setminus E \in \Sigma$, and disjointness of the A_i implies disjointness of the $A_i \setminus E$. By definition of E we have

$$s^*(x) = \begin{cases} 0 & \text{if } x \in E \\ f(x) & \text{if } x \in S \setminus E \end{cases} = \begin{cases} 0 & \text{if } x \in E \\ g(x) & \text{if } x \in S \setminus E \end{cases} \leq g(x).$$

Hence s^* is a simple function such that $0 \leq s^* \leq g$. Using that $m(E) = 0$, from (4.3) we can deduce that $\int_S s\,dm = \int_S s'\,dm$, which by (4.6) implies that $\int_S f\,dm \leq \int_S g\,dm$.

We can apply the same argument with the roles of f and g swapped to deduce that $\int_S g\,dm \leq \int_S f\,dm$, hence in fact they are equal. ■

We can apply the same idea to inequalities too. For example, we say that $f \leq g$ almost everywhere if $m(\{x \in \S - f(x) \leq g(x)\}) = 0$.

4.3 The monotone convergence theorem

It is helpful to work with non-negative measurable functions for a bit longer, before we take the final step of defining the Lebesgue integral. We continue to work over a general measure space (S, Σ, m).

In this section we will see interaction between integrals and limits for the first time. In particular we establish a set of conditions under which, for the Lebesgue integral, we can deduce that $\int_S f_n \, dm \to \int_S f \, dm$. This marks a key step in our development of the Lebesgue integral. It is the first point at which we see Lebesgue integration do something that Riemann integration cannot.

We say that a sequence (f_n) of functions is *monotone increasing* if $f_n \leq f_{n+1}$ for all $n \in \mathbb{N}$. Note that in this case by Lemma 2.1.1 the pointwise limit $f = \lim_{n\to\infty} f_n$ automatically exists, is non-negative and measurable, and may take values in $[0, \infty]$. Similarly, we say that (f_n) is *monotone decreasing* if $f_n \geq f_{n+1}$ for all n.

Theorem 4.3.1 (Monotone Convergence Theorem) *Let f_n, f be measurable functions from S to $[0, \infty)$. Suppose that:*

1. *for all $n \in \mathbb{N}$ we have $f_n \leq f_{n+1}$,*
2. *$f_n \to f$ almost everywhere.*

Then

$$\int_S f_n \, dm \to \int_S f \, dm \tag{4.7}$$

as $n \to \infty$.

Proof: Since (f_n) is increasing, $f^*(x) = \lim_{n\to\infty} f(x)$ exists for all $x \in \mathbb{R}$. By Lemma 3.1.5 we have that f^* is measurable. We have $f \overset{\text{a.e.}}{=} f^*$ by our second assumption, which by Lemma 4.2.6 means that $\int_s f \, dm = \int_S f^* \, dm$. Hence we may assume, without loss of generality by using f^* in place of f, that $f_n \to f$ pointwise.

We now aim to prove (4.7). Since (f_n) is increasing and pointwise convergent to f, we have $f = \sup_n f_n$. We have $f_1 \leq f_2 \leq \ldots \leq f$, so by monotonicity from Lemma 4.2.2 we have

$$\int_S f_1 \, dm \leq \int_S f_2 \, dm \leq \cdots \leq \int_S f \, dm.$$

Hence $n \mapsto \int_S f_n \, dm$ defines an increasing sequence that, by Lemma 2.1.1, has a limit in $[0, \infty]$ and

$$\lim_{n\to\infty} \int_S f_n \, dm \leq \int_S f \, dm. \tag{4.8}$$

In order to establish (4.7) we must also prove the reverse inequality. To simplify notation, let us write $a = \lim_n \int_S f_n dm$. So, we need to show that $a \geq \int_S f dm$. Let s be a simple function with $0 \leq s \leq f$ and choose $c \in \mathbb{R}$ with $0 < c < 1$. Our plan is to show that $a \geq c \int_S s \, dm$ and then take a sup over c and s.

For each $n \in \mathbb{N}$, let

$$E_n = \{x \in S; f_n(x) \geq cs(x)\}.$$

Note that $E_n \in \Sigma$ because by Theorem 3.1.5 the function $f_n - cs$ is measurable, and $E_n = (f_n - cs)^{-1}([0,\infty))$. We claim that

$$E_n \subseteq E_{n+1} \text{ for all } n \in \mathbb{N} \qquad \text{and} \qquad \bigcup_{n=1}^{\infty} E_n = S. \tag{4.9}$$

The first claim in (4.9) follows because (f_n) is increasing. To prove the second claim in (4.9), note first that if $x \in S$ with $s(x) = 0$ then $x \in E_n$ for all $n \in \mathbb{N}$. If $x \in S$ with $s(x) \neq 0$ then $f(x) \geq s(x) > cs(x)$, and since $f_n(x)$ is monotone increasing to $f(x)$ we must therefore have $N \in \mathbb{N}$ such that $f_n(x) \geq cs(x)$ for all $n \geq N$. For such n we have $x \in E_n$.

By parts 1 and 2 of Lemma 4.2.2 we have

$$a \geq \int_S f_n \, dm \geq \int_{E_n} f_n \, dm \geq \int_{E_n} cs \, dm. \tag{4.10}$$

The function s is simple, and it is easily checked that this means cs is also simple. Hence by Lemma 4.1.4 $\nu(X) = \int_X cs \, dm$ defines a measure. By (4.9) and Lemma 1.7.1 we have $\nu(E_n) \to \nu(S)$, that is $\int_{E_n} cs \, dm \to \int_S cs \, dm$ as $n \to \infty$. Letting $n \to \infty$ in (4.10) thus gives

$$a \geq \int_S cs \, dm.$$

Lemma 4.1.2 gives $\int_S cs \, dm = c \int_S s \, dm$, hence

$$a \geq c \int_S s \, dm.$$

This holds for any $c \in (0,1)$ and any simple function s with $0 \leq s \leq f$. Letting $c \uparrow 1$, and using that limits preserve weak inequalites, gives

$$a \geq \int_S s \, dm.$$

Taking a supremum over all simple functions s with $0 \leq s \leq f$ and using Definition 4.2.1 gives that

$$a \geq \int_S f \, dm.$$

This provides the reverse inequality to (4.8) and completes the proof. ■

We will look at examples of calculating integrals, using tools like the monotone convergence theorem, in Section 4.7. For now we press onwards with developing the Lebesgue integral, and upgrade the linearity property to cover non-negative measurable functions. Theorem 3.5.2 (which have not used until now!) turns out to be crucial here.

Lemma 4.3.2 *Let $f, g : S \to [0,\infty)$ be measurable. Then for all $\alpha, \beta \in [0,\infty)$ we have*

$$\int_S \alpha f + \beta g \, dm = \alpha \int_S f \, dm + \beta \int_S g \, dm.$$

PROOF: By Theorem 3.5.2 we can find an increasing sequence of simple functions (s_n) that converges pointwise to f and an increasing sequence of simple functions (t_n) that converges pointwise to g. Exercise **3.5** gives that $\alpha s_n + \beta t_n$ is a simple function, and it is clear that $\alpha s_n + \beta t_n$ is an increasing sequence of functions that converges pointwise to $\alpha f + \beta g$.

By Lemma 4.1.2 part 2 we have $\int_S \alpha s_n + \beta t_n \, dm = \alpha \int_S s_n \, dm + \beta \int_S t_n \, dm$. Letting $n \to \infty$ and using Theorem 4.3.1 to take the limit of all of the integral terms, we obtain that $\int_S \alpha f + \beta g \, dm = \alpha \int_S f \, dm + \beta \int_S g \, dm$ as required. ■

4.4 Integration as a measure

In Lemma 4.1.4 we saw that integrals of simple functions gave us a way of constructing measures. We'll now carry that property over to integrals of non-negative functions. In this case the upgrade to non-negative measurable functions provides the final version of the property. Other properties, such as monotonicity and linearity, will receive one more upgrade in Section 4.5. We continue to work over a general measure space (S, Σ, m).

Theorem 4.4.1 *Let $f : S \to [0, \infty)$ be measurable. Then $\nu : \Sigma \to [0, \infty]$ by*

$$\nu(A) = \int_A f\,dm$$

is a measure.

Proof: We check the two properties in Definition 3.1.1. We have $\nu(\emptyset) = \int_\emptyset f\,dm = \int_S \mathbb{1}_\emptyset f\,dm$. Since $\mathbb{1}_\emptyset = 0$ this gives $\nu(\emptyset) = \int_S 0\,dm$. The zero function is a simple function $0 = 0\mathbb{1}_S$, and (4.3) gives that it has integral zero. Thus $\nu(\emptyset) = 0$.

We need to show that ν is countably additive. Let $(E_n)_{n\in\mathbb{N}}$ be pairwise disjoint subsets of S and let $E = \bigcup_{n=1}^\infty E_n$. Set $F_n = \bigcup_{i=1}^n E_i$. Then $F_n \subseteq F_{n+1}$ and hence $\mathbb{1}_{F_n} \le \mathbb{1}_{F_{n+1}}$, so $\mathbb{1}_{F_n} f \le \mathbb{1}_{F_{n+1}} f$. Also, $\bigcup_{n=1}^\infty F_n = \bigcup_{n=1}^\infty E_n$, so $\mathbb{1}_{F_n} \to \mathbb{1}_E$ pointwise. Hence $\mathbb{1}_{F_n} f \to \mathbb{1}_E f$ pointwise so by Theorem 4.3.1 we have

$$\int_S \mathbb{1}_{F_n} f\,dm \to \int_S \mathbb{1}_E f\,dm. \tag{4.11}$$

The right hand side of the above is equal to $\int_E f\,dm = \nu(E)$. By Lemma 4.3.2 the left hand side is equal to

$$\int_S \sum_{i=1}^n \mathbb{1}_{E_i} f\,dm = \sum_{i=1}^n \int_S \mathbb{1}_{E_i} f\,dm = \sum_{i=1}^n \int_{E_i} f\,dm = \sum_{i=1}^n \nu(E_i).$$

Putting these into (4.11) gives that $\lim_n \sum_{i=1}^n \nu(E_i) = \nu(\bigcup_{i=1}^n E_i)$, as required. ■

Example 4.4.2 The *Gaussian measure* on $\mathbb{R}$ is obtained by taking $f = \phi$ where $\phi : \mathbb{R} \to \mathbb{R}$ is given by $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and taking m as Lebesgue measure. We note an explicit connection with probability theory:

$$I_A(\phi) = \int_A \frac{1}{2\pi} e^{-x^2/2}\,d\lambda(x)$$

which you should recognize as equal to $\mathbb{P}[Z \in A]$ where $Z \sim N(0, 1)$. Thus $A \mapsto \int_A \phi\,d\lambda$ is the law of a standard normal random variable. Normal random variables are often known as Gaussian random variables.

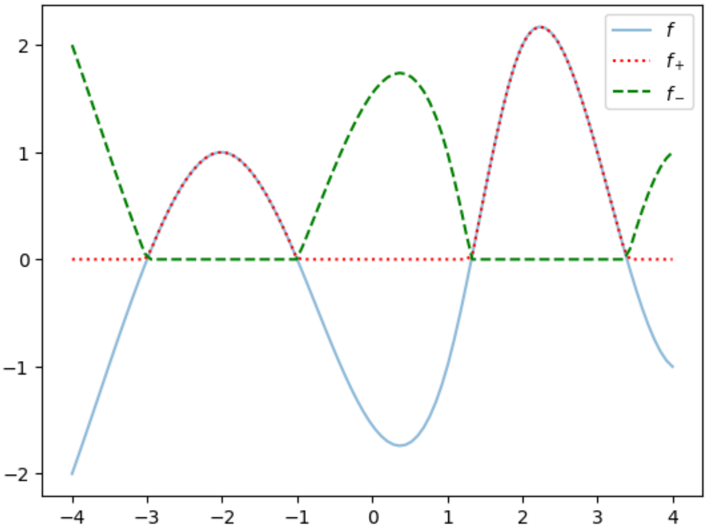
4.5 The Lebesgue integral

At last we are ready for the final step in the construction of the Lebesgue integral, which extends our current framework to a class of measurable functions that are real-valued, instead of just non-negative. This step requires a little preparation. We continue to work over an arbitrary measure space (S, Σ, m).

For a function $f : S \to \mathbb{R}$ we define

$$f_+ = \max(f, 0), \qquad f_- = \max(-f, 0),$$

which are pointwise definitions of functions $f_+, f_- : S \to \mathbb{R}$. It is easiest to see what is going on here with a picture:



Note that

$$f = f_+ - f_-,$$

and that by Theorem 3.1.5 both f_+ and f_- are non-negative measurable functions from S to $\mathbb{R}$. Step 2 (Definition 4.2.1) tells us the values of $\int_S f_+ \, dm$ and $\int_S f_- \, dm$, which are extended real numbers in $[0, \infty]$. Note also that $|f| = f_+ + f_-$, and that $\int_S |f| \, dm$ is also covered by Step 2.

Definition 4.5.1 (Lebesgue Integral, Step 3) Let $f : S \to \mathbb{R}$ be measurable. If *at least one* of $\int_S f_+ \, dm$ and $\int_S f_- \, dm$ is not equal to $+\infty$, then we define

$$\int_S f \, dm = \int_S f_+ \, dm - \int_S f_- \, dm, \tag{4.12}$$

which is an extended real number.

If both $\int_S f_+ \, dm$ and $\int_S f_- \, dm$ are equal to $+\infty$ then $\int_S f \, dm$ is undefined. Note that in this case (4.12) would give $\infty - \infty$, which is undefined.

It is important that we have a way to avoid handling infinities. This is provided by the following definition:

$$\mathcal{L}^1 = \left\{ f : S \to \mathbb{R} \,;\, f \text{ is measurable and } \int_S |f| \, dm < \infty \right\}. \tag{4.13}$$

It is common to refer to $f \in \mathcal{L}^1$ as *Lebesgue integrable* functions. We will not call them that within this course, because of potential confusion with Definition 4.5.1. We will mostly work with functions in $\mathcal{L}^1$ from now on. We might write $\mathcal{L}^1(S)$ or even $\mathcal{L}^1(S, \Sigma, m)$ if we need to be specific about which measure space we mean to use as the domain. The following lemma explains the connection between (4.12) and (4.13).

Lemma 4.5.2 *Let $f : S \to \mathbb{R}$ be measurable. Then $f \in \mathcal{L}^1$ if and only if both $\int_S f_+ \, dm$ and $\int_S f_- \, dm$ are finite.*

Proof: We have $0 \leq f_+ \leq |f|$ and $0 \leq f_- \leq |f|$, hence if $\int_S |f| \, dm < \infty$ then Lemma 4.2.2 gives that both $\int_S f_+ \, dm$ and $\int_S f - \, dm$ are finite. The reverse implication follows because $|f| = f_+ + f_-$, which are all non-negative measurable functions, hence by Lemma 4.3.2 we have $\int_S |f| \, dm = \int_S f_+ \, dm + \int_S f_- \, dm < \infty$. ■

We will begin to look at examples where we evaluate integrals in Section 4.7. First, let us complete the process of establishing the key properties of the Lebesgue integral.

Theorem 4.5.3 *Suppose that $f, g \in \mathcal{L}^1(S, \Sigma, m)$ and $A, B \in \Sigma$.*

1. *Domain additivity: If $A \cap B = \emptyset$ then*
$$\int_A f \, dm + \int_B f \, dm = \int_{A \cup B} f \, dm.$$
2. *Linearity: For all $\alpha, \beta \in \mathbb{R}$ we have $\alpha f + \beta g \in \mathcal{L}^1$ and*
$$\int_A \alpha f + \beta g \, dm = \alpha \int_A f \, dm + \beta \int_A g \, dm.$$
3. *Montonicity: If $f \leq g$ then*
$$\int_A f \, dm \leq \int_A g \, dm.$$
4. *Absolute values:*
$$\left| \int_A f \, dm \right| \leq \int_A |f| \, dm.$$
5. *Almost everywhere equality: If $f \overset{\text{a.e.}}{=} g$ then $\int_A f \, dm = \int_A g \, dm$.*

Proof: For non-negative measurable functions part 1 is contained within Theorem 4.4.1. Thus $\int_A f_\pm \, dm + \int_B f_\pm \, dm = \int_{A \cup B} f_\pm \, dm$. Subtracting the f_- case from the f_+ case gives

$$\int_A f_+ \, dm - \int_A f_- \, dm + \int_B f_+ \, dm - \int_B f_- \, dm = \int_{A \cup B} f_+ \, dm - \int_{A \cup B} f_- \, dm$$

which by (4.12) is exactly what we need to prove part 1.

For parts 2-5, it suffices to prove the case $S = A$. Note that in view of Definition 4.1.3 we can recover the general case by replacing f and g by $\mathbb{1}_A f$ and $\mathbb{1}_A g$. Part 5 follows immediately from Lemma 4.2.6 and (4.12). Several of the remaining parts are left for you as excerises. Part 3 is Exercise 4.4 part (c). Part 4 is Exercise 4.4 part (a). For part 2, Exercise 4.4 part (d) shows that

$$\alpha \int_S f \, dm = \int_S \alpha f \, dm \tag{4.14}$$

for all $f \in \mathcal{L}^1$. In order to prove part 2 it remains only to show that

$$\int_S f + g \, dm = \int_S f \, dm + \int_S g \, dm. \tag{4.15}$$

We will prove (4.15) here. Note that we may assume that both f, g are not identically 0, because $\int_S 0 \, dm = 0$. The fact that $f + g$ is in $\mathcal{L}^1$ if f and g are follows from Exercise 4.4 part (b).

To show (4.15) we first need to consider six different special cases. Writing $h = f + g$, these cases are (1) $f \geq 0, g \geq 0, h \geq 0$, (2) $f \leq 0, g \leq 0, h \leq 0$, (3) $f \geq 0, g \leq 0, h \geq 0$, (4) $f \leq 0, g \geq 0, h \geq 0$, (5) $f \geq 0, g \leq 0, h \leq 0$, (6) $f \leq 0, g \geq 0, h \leq 0$. Note that case (1) is precisely Lemma 4.3.2. We'll just prove case (3) here, and note that the others are similar. To show case (3) we write $f = h + (-g)$ and, noting that all these are non-negative functions, from Lemma 4.3.2 we obtain $\int_S f \, dm = \int_S (f + g) \, dm + \int_S (-g) \, dm$, and hence from (4.14) we have $\int_S (f + g) \, dm = \int_S f \, dm - \int_S (-g) \, dm = \int_S f \, dm + \int_S g \, dm$, which proves case (3).

For a general $f \in \mathcal{L}^1$, we write $S = S_1 \cup S_2 \cup S_3 \cup S_4 \cup S_5 \cup S_6$, where S_l is the set of all $x \in S$ for which case (l) holds for $l = 1, 2, \ldots, 6$. These sets are disjoint and measurable. We then have

$$\int_S (f + g) \, dm = \sum_{i=1}^{6} \int_{S_i} (f + g) \, dm = \sum_{i=1}^{6} \int_{S_i} f \, dm + \sum_{i=1}^{6} \int_{S_i} g \, dm = \int_S f \, dm + \int_S g \, dm.$$

In the above, the first and last equalities use part 1 of the present lemma, and the middle equality uses cases (1)-(6) above. This completes the proof of part 2. ■

4.6 The dominated convergence theorem

We continue to work over a general measure space (S, Σ, m). Let $f, g : S \to \mathbb{R}$ be measurable. We say that g *dominates* f if $|f| \leq |g|$ almost everywhere. This concept is naturally connected to $\mathcal{L}^1$ via the following lemma.

Lemma 4.6.1 *Let $g \in \mathcal{L}^1$ and suppose that $f : S \to \mathbb{R}$ is measurable with $|f| \leq |g|$ almost everywhere. Then $f \in \mathcal{L}^1$.*

PROOF: By part 1 of Lemma 4.2.2 we have $\int_S |f|\,dm \leq \int_S |g|\,dm$, which completes the proof. (Exercise: Why don't we use Theorem 4.5.3 here?) ■

We now present the second of our convergence theorems, the famous *Lebesgue dominated convergence theorem* - an extremely powerful tool in both the theory and applications of modern analysis.

Theorem 4.6.2 (Dominated Convergence Theorem) *Let f_n, f be functions from S to $\mathbb{R}$. Suppose that f_n is measurable and:*

1. *There is a function $g \in \mathcal{L}^1$ such that $|f_n| \leq |g|$ almost everywhere.*
2. *$f_n \to f$ almost everywhere.*

Then $f \in \mathcal{L}^1$ and

$$\int_S f_n dm \to \int_S f dm$$

as $n \to \infty$.

The monotone convergence theorem is the basis of the interaction of Lebesgue integrals with limits, but it has the disadvantage that it only applies to monotone sequences of functions. The dominated convergence theorem does not require monotonicity but does require that all functions involved are dominated by some $g \in \mathcal{L}^1$. For this reason g is often known as a *dominating function* for the (f_n). The monotone and dominated convergence theorems are often known for short as the MCT and DCT. The proof of the DCT appears in Section 4.6.1. This completes our development of the Lebesgue integral.

We're now in a position to use Lebesgue integration for doing calculations with integrals. We'll do so in Section 4.7, which will include examples of both the MCT and DCT.

4.6.1 Proof of the DCT ($\star$)

The proof of the DCT is off-syllabus, although we might cover it in lectures if we have time. We'll begin with a famous lemma before we give the main proof. It is called *Fatou's lemma* after the French mathematician and astronomer Pierre Fatou (1878-1929). It tells us how $\liminf$ and $\int$ interact. Understanding this is they key step for moving from monotone seqences of functions; non-monotone sequences of functions might not have pointwise limits but they do always have pointwise $\liminf$s.

Lemma 4.6.3 (Fatou's Lemma) *If (f_n) is a sequence of non-negative measurable functions from S to $\mathbb{R}$ then*

$$\liminf_{n\to\infty}\int_S f_n\,dm \geq \int_S \liminf_{n\to\infty} f_n\,dm$$

PROOF: Define $g_n = \inf_{k\geq n} f_k$. Then (g_n) is an increasing sequence which converges to $\liminf_{n\to\infty} f_n$. Now as $f_l \geq \inf_{k\geq n} f_k$ for all $l \geq n$, Lemma 4.2.2(1)) we have that for all $l \geq n$

$$\int_S f_l\,dm \geq \int_S \inf_{k\geq n} f_k\,dm,$$

and so

$$\inf_{l\geq n}\int_S f_l\,dm \geq \int_S \inf_{k\geq n} f_k\,dm.$$

Take limits on both sides of this last inequality and then apply the monotone convergence theorem (on the right hand side) to obtain

$$\begin{aligned}\liminf_{n\to\infty}\int_S f_n\,dm &\geq \lim_{n\to\infty}\int_S \inf_{k\geq n} f_k\,dm\\ &= \int_S \lim_{n\to\infty}\inf_{k\geq n} f_k\,dm\\ &= \int_S \liminf_{n\to\infty} f_n\,dm\end{aligned}$$

as required. ■

Note that we do not require (f_n) to be a bounded sequence, so $\liminf_{n\to\infty} f_n$ should be interpreted as an extended measurable function, as discussed at the end of Chapter 2. The corresponding result for $\limsup$, in which case the inequality is reversed, is known as the *reverse Fatou lemma* and can be found as Exercise **4.12**.

PROOF OF THEOREM 4.6.2: Note that we didn't assume explicitly that $f_n \in \mathcal{L}^1$, because this fact follows immediately from the first assumption and Lemma 4.6.1. For the same reason as in the proof of Theorem 4.3.1, we may assume without loss of generality that $f_n \to f$ pointwise and that $|f_n| \leq |g|$ pointwise. Thus f is measurable by Theorem 3.1.5. We may further assume without loss of generality that $|f_n| \leq g$, by using $|g|$ in place of g and noting that $g \in \mathcal{L}^1 \Leftrightarrow |g| \in \mathcal{L}^1$.

Since $f_n \to f$ pointwise, $|f_n| \to |f|$ pointwise. By Fatou's lemma (Lemma 4.6.3) and monotonicity from Theorem 4.5.3, we have

$$\begin{aligned}\int_S |f|\,dm &= \int_S \liminf_{n\to\infty} |f_n|\,dm\\ &\leq \liminf_{n\to\infty}\int_S |f_n|\,dm\end{aligned}$$

$$\leq \int_S g\,dm < \infty,$$

so $f \in \mathcal{L}^1$.

For all $n \in \mathbb{N}$, since $|f_n| \leq g$ we have $g + f_n \geq 0$, so by Fatou's lemma

$$\int_S \liminf_{n\to\infty}(g + f_n)\,dm \leq \liminf_{n\to\infty} \int_S (g + f_n)\,dm. \tag{4.16}$$

As pointwise limits we have $\liminf_n(g + f_n) = g + \lim_n f_n = g + f$ so by linearity from Theorem 4.5.3 we have

$$\begin{aligned}
&\int_S \liminf_{n\to\infty}(g + f_n)\,dm = \int_S g + \liminf_{n\to\infty} f_n\,dm = \int_S g\,dm + \int_S \liminf_{n\to\infty} f_n\,dm, \\
&\liminf_{n\to\infty} \int_S (g + f_n)\,dm = \int_S g\,dm + \liminf_{n\to\infty} \int_S f_n\,dm
\end{aligned}$$

Putting these two equations in (4.16) gives that

$$\int_S f dm \leq \liminf_{n\to\infty} \int_S f_n dm. \tag{4.17}$$

Next, we repeat the argument with $g + f_n$ replaced by $g - f_n$. Note that $|f_n| \leq g$ gives that $g - f_n$ is also non-negative for all $n \in \mathbb{N}$. The result of doing so is

$$-\int_S f dm \leq \liminf_{n\to\infty} \left(-\int_S f_n dm\right)$$

which from Lemma 2.2.3 rearranges to

$$\int_S f dm \geq \limsup_{n\to\infty} \int_S f_n dm \tag{4.18}$$

Combining (4.17) and (4.18) we see that

$$\limsup_{n\to\infty} \int_S f_n dm \leq \int_S f dm \leq \liminf_{n\to\infty} \int_S f_n dm. \tag{4.19}$$

Recall that $\liminf_n a_n \leq \limsup_n a_n$ for any sequence, so in fact all three terms in (4.19) are equal. It now follows from Lemma 2.2.2 that $\int_S f dm = \lim_n \int_S f_n dm$. ■

4.7 Calculations with the Lebesgue integral

From now on, we'll often write $\int_S f(x)\,dm(x)$ in place of $\int_S f\,dm$. This notation has the advantage that we can write an expression like $\int_S e^{-x^2}\,d\lambda(x)$ without having to specify the function $f(x) = e^{-x^2}$ in advance. We follow common convention and write $\int \ldots dx$ for integration with respect to Lebesgue measure, which we would formally write as $\int \ldots d\lambda(x)$. We'll allow ourselves to do this in cases where its clear from the context that we mean to integrate with respect to Lebesgue measure.

Remark 4.7.1 ($\star$) In some countries (e.g. France) it is common to write $\int_S dm\,f$ instead of $\int_S f\,dm$. We won't use that notation within this course. The notation $\int_S f(x)m(dx)$ in place of $\int_S f(x)dm(x)$ is also widely used.

We will show in Theorem 4.10.3 that, if $f : [a,b] \to \mathbb{R}$ is Riemann integrable, over a closed bounded interval $[a,b] \subseteq \mathbb{R}$, then f is also Lebesgue integrable (i.e $f \in \mathcal{L}^1$), and in this case the value of the two integrals is equal. Consequently, you may use all the facts you already know about Riemann integration on $\mathbb{R}$ to evaluate integrals of the form $\int_{[a,b]} f(x)\,dx$, which we would normally write as $\int_a^b f(x)\,dx$. This includes integration by substitution, by parts, the Fundamental Theorem of Calculus, and so on. A word of warning is necessary: such results only apply on *bounded* intervals $[a,b]$ and in general are not true on unbounded intervals.

The Lebesgue integral also gives us the MCT and DCT, for when we need limits and integrals to interact. In this section we look at a few examples of using the major results of Chapter 4 to integrate particular functions. We'll begin with the testing for $\mathcal{L}^1$ and using the monotone convergence theorem, before moving on to the dominated convergence theorem.

Example 4.7.2 We aim show that $f(x) = x^{-\alpha}$ is in $\mathcal{L}^1$ on $[1,\infty)$ for $\alpha > 1$.

For each $n \in \mathbb{N}$ define $f_n(x) = x^{-\alpha}\mathbb{1}_{[1,n]}(x)$. Then $(f_n(x))$ increases to $f(x)$ as $n \to \infty$. We have

$$\int_1^\infty f_n(x)dx = \int_1^n x^{-\alpha}dx = \frac{1}{\alpha-1}(1-n^{1-\alpha}).$$

By the monotone convergence theorem,

$$\int_1^\infty x^{-\alpha}dx = \frac{1}{\alpha-1}\lim_{n\to\infty}(1-n^{1-\alpha}) = \frac{1}{\alpha-1}.$$

Example 4.7.3 We aim to show that $f(x) = x^\alpha e^{-x}$ is in $\mathcal{L}^1$ on $[0,\infty)$ for $\alpha > 0$.

The key idea here is that e^{-x} tends to zero very quickly as $x \to \infty$, and we can use this fast convergence to overpower the 'opposing' fact that $x^\alpha \to \infty$. Recall that for any $M \geq 0$ we have $\lim_{x\to\infty} x^M e^{-x} = 0$, so that given any $\epsilon > 0$ there exists $R > 0$ so that $x > R \Rightarrow x^M e^{-x} < \epsilon$, and choose M so that $M - \alpha > 1$. Now write

$$x^\alpha e^{-x} = x^\alpha e^{-x}\mathbb{1}_{[0,R]}(x) + x^\alpha e^{-x}\mathbb{1}_{(R,\infty)}(x).$$

By part (b) of Exercise 4.4 the sum of two $\mathcal{L}^1$ functions is in $\mathcal{L}^1$, so we'll aim to prove that both terms on the right hand side are in $\mathcal{L}^1$. The first term on the right clearly is, because it is bounded on $[0,R]$ and zero elsewhere. For the second term we use that fact that for all $x > R$,

$$x^\alpha e^{-x} = x^M e^{-x}.x^{\alpha-M} < \epsilon x^{\alpha-M},$$

and which is thus in $\mathcal{L}^1$ by Example 4.7.2. So the result follows by Lemma 4.6.1.

Example 4.7.4 We want to calculate

$$\lim_{n\to\infty}\int_0^1 \frac{nx^2}{nx+5}\,dx.$$

We'll work in the measure space $([0,1],\mathcal{B}([0,1]),\lambda)$ and consider the sequence of functions (f_n) where $f_n(x)=\frac{nx^2}{nx+5}$ for all $x\in[0,1], n\in\mathbb{N}$. Each f_n is continuous, hence also measurable by Lemma 3.2.1. It is straightforward to check that $\lim_{n\to\infty} f_n(x)=x$ for all $x\in[0,1]$ and that $|f_n(x)|\leq 1$ for all $n\in\mathbb{N}, x\in[0,1]$. So in this case, we can take $K=1$, and apply the DCT with dominating function $g=1$ to deduce that $f(x)=x$ is in $\mathcal{L}^1$, and

$$\lim_{n\to\infty}\int_{[0,1]}\frac{nx^2}{nx+5}\,dx=\int_{[0,1]} x\,dx.$$

We finish by evaluating $\int_{[0,1]} x\,dx=\left[\frac{x^2}{2}\right]_0^1=\frac{1}{2}$.

Example 4.7.5 Summation of series is a special case of Lebesgue integration. Suppose that we are interested in $\sum_{n=1}^\infty a_n$, where $a_n\geq 0$ for all $n\in\mathbb{N}$. We consider the sequence (a_n) as a function $a:\mathbb{N}\to[0,\infty)$. We work with the measure space $(\mathbb{N},\mathcal{P}(\mathbb{N}),m)$ where m is counting measure. Then every sequence (a_n) gives rise to a non-negative measurable function a and

$$\sum_{n=1}^\infty a_n=\int_{\mathbb{N}} a(n)\,dm(n),$$

which is for you to show in Problem 4.10. The same formula holds for general $(a_n)\subseteq\mathbb{R}$ provided that $\sum_n|a_n|<\infty$ i.e. that $a\in\mathcal{L}^1(\mathbb{N})$.

In this context the monotone and dominated converge theorems provide tools for working with *sequences of series*. For example, suppose that for each $m\in\mathbb{N}$ we have a sequence $(a_m(n))_{n\in\mathbb{N}}$, given by

$$a_m(n)=\frac{1}{n^3}+\frac{1}{1+m^2n^2}.$$

We can't easily compute the value of $\sum_{n\in\mathbb{N}} a_m(n)$ for any given m. But we can note that $a_m(n)\to\frac{1}{n^3}$ as $m\to\infty$, for all n, and that $|a_m(n)|\leq g(n)=\frac{1}{n^3}+\frac{1}{n^2}$. We know from analysis that $\sum_n\frac{1}{n^3}$ and $\sum_n\frac{1}{n^2}$ are both finite, so $g\in\mathcal{L}^1(\mathbb{N},\mathcal{P}(\mathbb{N}),m)$ – which in this setting is just the claim that $\sum_n g(n)<\infty$. So the Dominated Convergence Theorem applies, and we obtain

$$\lim_{m\to\infty}\sum_{n\in\mathbb{N}}\left(\frac{1}{n^3}+\frac{1}{1+m^2n^2}\right)=\sum_{n\in\mathbb{N}}\frac{1}{n^3}.$$

4.8 Lebesgue integration of complex valued functions (Δ)

The definition of the Lebesgue integral can be extended to complex valued functions. Let (S, Σ, m) be a measure space and $f : S \to \mathbb{C}$. We can always write $f = f_1 + if_2$, where the real and imaginary parts are $f_i : S \to \mathbb{R}$ $(i = 1, 2)$. If both f_1 and f_2 are measurable then we say that f is measurable[1]. If both f_1 and f_2 have integrals according to Definition 4.5.1 then we define

$$\int_S f\,dm = \int_S f_1\,dm + i\int_S f_2\,dm.$$

Recall that for $z = x + iy \in \mathbb{C}$, where x and y are the real and imaginary parts of f, we define $|z| = (x^2 + y^2)^{1/2}$. Correspondingly, we make the pointwise definition that when $f : S \to \mathbb{C}$, $|f| = (f_1 + f_2)^{1/2}$. Note that $|f| : S \to \mathbb{R}$. This allows us to define a complex version of $\mathcal{L}^1$, given by

$$\mathcal{L}^1_{\mathbb{C}} = \left\{ f : S \to \mathbb{C}\,;\, f \text{ is measurable and } \int_S |f|\,dm < \infty \right\}. \tag{4.20}$$

When we need to specify that we mean the complex/real versions of $\mathcal{L}^1$ we will write $\mathcal{L}^1_{\mathbb{C}}$ and $\mathcal{L}^1_{\mathbb{R}}$. When it is clear which one we mean we will simply write $\mathcal{L}^1$. The extension of Lemma 4.5.2 is as follows.

Lemma 4.8.1 *Let $f : S \to \mathbb{C}$ be measurable, with real part $f_1 : S \to \mathbb{R}$ and imaginary part $f_2 : S \to \mathbb{R}$. Then the following are equivalent:*

1. $f \in \mathcal{L}^1_{\mathbb{C}}$,
2. $|f| \in \mathcal{L}^1_{\mathbb{R}}$,
3. $f_1 \in \mathcal{L}^1_{\mathbb{R}}$ *and* $f_2 \in \mathcal{L}^1_{\mathbb{R}}$.

PROOF: Parts 1 and 2 are equivalent by (4.13) and (4.20). Parts 2 and 3 are equivalent by Lemma 4.5.2. ■

All of our results so far may be applied to the real and imaginary parts of a complex valued function $f : S \to \mathbb{C}$, or to its absolutely value $|f|$. It is often more convenient when we can apply them directly, so let us make some notes on this.

- There is no 'less than or equal to' for complex numbers. This means that some results have no natural equivalents in $\mathbb{C}$, essentially anything involving $\leq$, min, max, inf, sup, lim inf or lim sup. For example, there is no monotonicity of integrals and no monotone convergence theorem.
- Several results continue to work without any modification, with the understanding that we must interpret $|\cdot|$ as the complex modulus and use $\mathcal{L}^1_{\mathbb{C}}$ in place of $\mathcal{L}^1_{\mathbb{R}}$. Most importantly, Lemma 4.6.1 and the dominated convergence theorem continue to hold, as do all of the properties in Theorem 4.5.3 except monotonicity.

  In such cases, the results for $\mathbb{C}$ can be proved by applying the result for $\mathbb{R}$ to both real and imaginary parts.
- Theorem 4.4.1 (integration as a measure) has no equivalent in $\mathbb{C}$, because measures must have real values[2].

[1] $(\star)$ Following on from Remark 3.3.4, it can be shown that this definition is equivalent to asking that $f : S \to \mathbb{C}$ be measurable with respect to $\mathcal{B}(\mathbb{C})$, where $\mathcal{B}(\mathbb{C})$ is generated by the open sets of the metric space $(\mathbb{C}, d)$ where $d(z, w) = |z - w|$.

[2] $(\star)$ In fact, there is a theory of complex and even vector valued measures, but it is outside of what we can cover.

4.9 Multiple integrals and function spaces ($\star$)

This section is included for interest. It is marked with a ($\star$) and it is off-syllabus. It includes two separate topics.

4.9.1 Fubini's Theorem ($\star$)

Fubini and Tonelli's theorems let us deal with expressions involving multiple integrals, for example of the form $\int_{S_1}\int_{S_2} f(x,y)\,dx\,dy$. They give different sets of conditions under which we can change the order of integration. Loosely, Tonelli's theorem is an analogue of the MCT and Fubini's theorem is an analogue of the DCT. Before introducing them we need to handle some measure theoretic details.

For $i=1,2$ let (S_i,Σ_i,m_i) be measure spaces and recall the product measure space $(S_1\times S_2,\Sigma_1\otimes\Sigma_2,m_1\times m_2)$ introduced in Section 1.9. If $f:S_1\times S_2\to\mathbb{R}$ is measurable then the coordinate projections $x\mapsto f(x,y)$ and $y\mapsto f(x,y)$ are measurable for almost all $x\in S_1$ and $y\in S_2$. Moreover, if $f\in\mathcal{L}^1(S_1\times S_2)$ then these coordinate projections are respectively in $\mathcal{L}^(S_1)$ and $\mathcal{L}^1(S_2)$, again for almost all $x\in S_1$ and $y\in S_2$, and the same is true of the functions $x\to\int_{S_2} f(x,y)m_2(dy)$ and $y\to\int_{S_2} f(x,y)m_2(dx)$. We won't include a proof of these claims here. They put us a in a position to state the key result of this section.

Theorem 4.9.1 *Let $f:S_1\times S_2\to\mathbb{R}$. Suppose that at least one of the following conditions holds.*

1. *Fubini's Theorem: $f\in\mathcal{L}^1(S_1\times S_2)$.*
2. *Tonelli's Theorem: $f\geq 0$.*

Then

$$\begin{aligned}\int_{S_1\times S_2} f\,d(m_1\times m_2) &= \int_{S_1}\left(\int_{S_2} f(x,y)\,dm_2(y)\right)dm_1(x)\\ &= \int_{S_2}\left(\int_{S_1} f(x,y)\,dm_1(x)\right)dm_2(y).\end{aligned}$$

In the case of Fubini's theorem $\int_{S_1\times S_2} f\,d(m_1\times m_2)$ is a real number, whilst in the case of Tonelli's theorem it is in $[0,\infty]$. These results are named after the Italian mathematicians Guido Fubini (1879-1943) and Leonida Tonelli (1885-1946). They are very important results, equal in stature to the MCT and DCT, but we omit a full treatment of them from our course in order to progress on to thinking about probability.

4.9.2 Function Spaces ($\star$)

This section is aimed at those taking courses in functional analysis. An important application of Lebesgue integration is to the construction of Banach spaces $\mathcal{L}^p(S,\Sigma,m)$ of equivalence classes of real-valued functions, under the equivalence relation $f\overset{\text{a.e.}}{=}g$, which satisfy the requirement

$$||f||_p=\left(\int_S |f|^p\,dm\right)^{\frac{1}{p}}<\infty,$$

where $1\leq p<\infty$. The function $||\cdot||_p$ is a norm on $L^p(S,\Sigma,m)$ if $p\geq 1$, but it is not a norm for $p<1$. This is the reason why, in Section 7.2, we will only define $\mathcal{L}^p$ convergence for $p\geq 1$.

When $p = 2$ we obtain a Hilbert space with inner product:

$$\langle f, g\rangle = \int_S fg\,dm.$$

There is also a Banach space $L^\infty(S, \Sigma, m)$ where

$$||f||_\infty = \inf\{M \geq 0; |f(x)| \leq M \text{ a.e.}\}.$$

Variants of all of these spaces exist with $\mathbb{C}$ in place of $\mathbb{R}$, using Lebesgue integration over $\mathbb{C}$ as defined in Section 4.8. These spaces play important roles in functional analysis.

4.10 Riemann integration ($\star$)

In this section, our aim is to show that if a bounded function $f : [a, b] \to \mathbb{R}$ is Riemann integrable, then it is measurable and Lebesgue integrable. Moreover, in this case the Riemann and Lebesgue integrals of f are equal. We state this result formally as Theorem 4.10.3. In this section we will prefer to say 'f is Lebesgue integral' rather than $f \in \mathcal{L}^1$, simply because it makes for better grammar when we compare Riemann and Lebesgue integrability within the same sentence.

We begin by briefly revising the Riemann integral. Note that this whole section is marked with a ($\star$), meaning that it is off-syllabus. It will be discussed briefly in lectures.

4.10.1 The Riemann integral ($\star$)

A partition $\mathcal{P}$ of $[a, b]$ is a set of points $\{x_0, x_1, \ldots, x_n\}$ with $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$. Define $m_j = \inf_{x_{j-1}\le x\le x_j} f(x)$ and $M_j = \sup_{x_{j-1}\le x\le x_j} f(x)$. We underestimate by defining

$$L(f, \mathcal{P}) = \sum_{j=1}^{n} m_j(x_j - x_{j-1}),$$

and overestimate by defining

$$U(f, \mathcal{P}) = \sum_{j=1}^{n} M_j(x_j - x_{j-1}),$$

A partition $\mathcal{P}'$ is said to be a *refinement* of $\mathcal{P}$ if $\mathcal{P} \subset \mathcal{P}'$. We then have

$$L(f, \mathcal{P}) \le L(f, \mathcal{P}'), \quad U(f, \mathcal{P}') \le U(f, \mathcal{P}). \tag{4.21}$$

A sequence of partitions $(\mathcal{P}_n)$ is said to be *increasing* if $\mathcal{P}_{n+1}$ is a refinement of $\mathcal{P}_n$ for all $n \in \mathbb{N}$.

Now define the *lower integral* $L_{a,b}f = \sup_{\mathcal{P}} L(f, \mathcal{P})$, and the *upper integral* $U_{a,b}f = \inf_{\mathcal{P}} U(f, \mathcal{P})$. We say that f is *Riemann integrable* over $[a, b]$ if $L_{a,b}f = U_{a,b}f$, and we then write the common value as $\int_a^b f(x)dx$. In particular, every continuous function on $[a, b]$ is Riemann integrable. The next result is very useful:

Theorem 4.10.1 *The bounded function f is Riemann integrable on $[a, b]$ if and only if for every $\epsilon > 0$ there exists a partition $\mathcal{P}$ for which*

$$U(f, \mathcal{P}) - L(f, \mathcal{P}) < \epsilon. \tag{4.22}$$

If (4.22) holds for some $\mathcal{P}$, it also holds for all refinements of $\mathcal{P}$. A useful corollary is:

Corollary 4.10.2 *If the bounded function f is Riemann integrable on $[a, b]$, then there exists an increasing sequence $(\mathcal{P}_n)$ of partitions of $[a, b]$ for which*

$$\lim_{n\to\infty} U(f, \mathcal{P}_n) = \lim_{n\to\infty} L(f, \mathcal{P}_n) = \int_a^b f(x)dx$$

Proof: This follows from Theorem (4.10.1) by successively choosing $\epsilon = 1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}, \ldots$. If the sequence $(\mathcal{P}_n)$ is not increasing, then just replace $\mathcal{P}_n$ with $\mathcal{P}_n \cup \mathcal{P}_{n-1}$ and observe that this can only improve the inequality (4.22). ∎

We can integrate *many* more functions using Lebesgue integration than we could using Riemann integration. For example, with Riemann integration we could not conclude that $\int_{[a,b]} \mathbb{1}_{\mathbb{R}\setminus\mathbb{Q}}(x)dx = (b - a)$, but with Lebesgue integration we can.

4.10.2 The connection ($\star$)

Theorem 4.10.3 *If $f : [a, b] \to \mathbb{R}$ is Riemann integrable, then it is Lebesgue integrable, and the two integrals coincide.*

PROOF: We use the notation λ for Lebsgue measure in this section. We also write $M = \sup_{x\in[a,b]} |f(x)|$ and $m = \inf_{x\in[a,b]} |f(x)|$.

Let $\mathcal{P}$ be a partition as above and define simple functions,

$$g_{\mathcal{P}} = \sum_{j=1}^{n} m_j \mathbb{1}_{(x_{j-1},x_j]}, \quad h_{\mathcal{P}} = \sum_{j=1}^{n} M_j \mathbb{1}_{(x_{j-1},x_j]}.$$

Consider the sequences (g_n) and (h_n) which correspond to the partitions of Corollary 4.10.2 and note that

$$L_n(f) = \int_{[a,b]} g_n d\lambda, \quad U_n f = \int_{[a,b]} h_n d\lambda,$$

where $U_n(f) = U(f, \mathcal{P}_n)$ and $L_n(f) = L(f, \mathcal{P}_n)$. Clearly we also have for each $n \in \mathbb{N}$,

$$g_n \leq f \leq h_n. \tag{4.23}$$

Since (g_n) is increasing (by (4.21)) and bounded above by M, it converges pointwise to a measurable function g. Similarly (h_n) is decreasing and bounded below by m, so it converges pointwise to a measurable function h. By (4.23) we have

$$g \leq f \leq h. \tag{4.24}$$

Again since $\max_{n\in\mathbb{N}}\{|g_n|, |h_n|\} \leq M$, we can use dominated convergence to deduce that g and h are both integrable on $[a, b]$ and by Corollary 4.10.2,

$$\int_{[a,b]} g d\lambda = \lim_{n\to\infty} L_n(f) = \int_a^b f(x)dx = \lim_{n\to\infty} U_n(f) = \int_{[a,b]} h d\lambda.$$

Hence we have

$$\int_{[a,b]} (h - g) d\lambda = 0,$$

and so by Corollary 3.3.1, $h(x) = g(x)$ (a.e.). Then by (4.24) $f = g$ (a.e.) and so f is measurable[3] and also integrable. So $\int_{[a,b]} f d\lambda = \int_{[a,b]} g d\lambda$, and hence we have

$$\int_{[a,b]} f d\lambda = \int_a^b f(x)dx.$$

■

[3] I'm glossing over a subtlety here. It is not true in general, that a function that is almost everywhere equal to a measurable function is measurable. It works in this case due to a special property of the Borel σ-field known as 'completeness'.

4.10.3 Discussion ($\star$)

An important caveat is that Theorem 4.10.3 only applies to *bounded* closed intervals. On unbounded intervals, there are examples of functions are Riemann integrable[4] but not Lebesgue integrable. One such example is $\int_0^\infty \frac{\sin x}{x}\,dx$. The function $\frac{\sin x}{x}$ oscillates above and below 0 as $x \to \infty$, and the Riemann integral $\int_0^\infty \frac{\sin x}{x}\,dx = \lim_{X\to\infty} \int_0^X \frac{\sin x}{x}\,dx$ only exists because these oscillations cancel each other out. In Lebesgue integration this isn't allowed to happen, and $\frac{\sin x}{x}$ fails to be Lebesgue integrable because $\int_0^\infty |\frac{\sin x}{x}|\,dx = \infty$. In fact, $\int_0^\infty (\frac{\sin x}{x})_-\,dx = \int_0^\infty (\frac{\sin x}{x})_+\,dx = \infty$, so $\int_0^\infty |\frac{\sin x}{x}|\,dx$ is undefined for the Lebesgue integral.

Let's discuss these ideas in the context of infinite series which, as we showed in Example 4.7.5, are a special case of the Lebesgue integral. That is,

$$\int_{\mathbb{N}} a_n\,d\#(n) = \sum_{n=1}^\infty a_n$$

where $a : \mathbb{N} \to \mathbb{R}$ is a sequence, and $\#$ is the counting measure on $\mathbb{N}$. Note that (a_n) is integrable if and only if $\sum_n |a_n| < \infty$, which is usually referred to as 'absolute convergence' in the context of infinite series. The key is that when infinite series are absolutely convergent they are much better behaved, as the following result shows. A 're-ordering' of a series simply means arranging its terms in a different order.

Theorem 4.10.4 *Let (a_n) be a real sequence.*

1. *Suppose $\sum_{n=1}^\infty |a_n| = \infty$ and $a_n \to 0$. Then, for any $\alpha \in \mathbb{R}$, there is a re-ordering $b_n = a_{p(n)}$ such that $\sum_{i=1}^n b_n \to \alpha$.*
2. *Suppose $\sum_n |a_n| < \infty$. Then, for any re-ordering $b_n = a_{p(n)}$, we have $\sum_{n=1}^\infty a_n = \sum_n b_n \in \mathbb{R}$.*

Imagine if we allowed something similar to case 1 was allowed to happen in integration, and let us think about integration over $\mathbb{R}$. It would mean that re-ordering the x-axis values (e.g. swap $[0,1)$ with $[1,2)$ and so on) could change the value of $\int_{\mathbb{R}} f(x)\,dx$! This would be nonsensical, and mean that integration over $\mathbb{R}$ no longer had anything to do with 'area under the curve'. So we have to avoid it, and we do so by restricting to integrable functions. Only then can we find nice conditions for 'limit' theorems like the dominated convergence theorem. Lebesgue integration solves this problem; Riemann integration cannot.

[4] Strictly, we should say 'improperly' Riemann integrable.

4.11 Exercises on Chapter 4

On integrals of simple functions and non-negative functions

4.1 Let $f : \mathbb{R} \to \mathbb{R}$ be defined as follows

$$f(x) = \begin{cases} 2 & \text{if } x \in [-2,-1] \\ -1 & \text{if } x \in (-1,1) \\ 3 & \text{if } x \in [1,2) \\ -5 & \text{if } x \in [2,3). \end{cases}$$

(a) Write f explicitly as a simple function and calculate $\int_{\mathbb{R}} f(x)\,dx$.

(b) Write down f_+ and f_- and confirm that they are non-negative simple functions. Calculate $\int_{\mathbb{R}} f_+(x)\,dx$ and $\int_{\mathbb{R}} f_-(x)\,dx$ and check that $\int_{\mathbb{R}} f_+(x)\,dx - \int_{\mathbb{R}} f_-(x)\,dx = \int_{\mathbb{R}} f(x)\,dx$.

4.2 Let (S, Σ, m) be a measure space, $A \in \Sigma$ and $f : S \to \mathbb{R}$ be a simple function. Show that $\mathbb{1}_A f$ is also a simple function.

4.3 Use Lemma 4.2.3 (Markov's inequality) to prove the following version of *Chebychev's inequality*. If $f : S \to \mathbb{R}$ is a measurable function and $c > 0$ then

$$m(\{x \in S; |f(x)| \geq c\}) \leq \frac{1}{c^2} \int_S f^2\,dm.$$

Formulate and prove a similar inequality where c^2 is replaced by c^p for $p \geq 1$.

4.4 This question contributes to the proof of Theorem 4.5.3. Let $f, g \in L^1(S, \Sigma, m)$ and let $\alpha \in \mathbb{R}$. Note that the integrals of f and g are defined by (4.12). Using the properties of integrals of non-negative functions that were proved in Sections 4.2 and 4.3, show that:

(a) $|\int_S f\,dm| \leq \int_S |f|\,dm$

(b) $\int_S |f+g|\,dm \leq \int_S |f|\,dm + \int_S |g|\,dm$

(c) $\alpha \int_S f\,dm = \int_S \alpha f\,dm$

Hint: First consider $\alpha \geq 0$*, then* $\alpha = -1$ *and then combine to handle* $\alpha < 0$.

(d) If $f \leq g$ then $\int_S f\,dm \leq \int_S g\,dm$.

In part (d) you may use linearity (of the full Lebesgue integral) because parts (b) and (c) provide the missing pieces that complete the proof of linearity.

On $\mathcal{L}^1$ and the convergence theorems

4.5 Let (S, Σ, m) be a measure space and suppose that m is a finite measure. Suppose that $f : S \to \mathbb{R}$ is bounded. Show that $f \in \mathcal{L}^1$.

4.6 Determine if the function $g : (0,1) \to \mathbb{R}$ by $g(x) = \log x$ is in $\mathcal{L}^1$.

4.7 Consider the sequence (f_n) on the measure space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ where $f_n = n\mathbb{1}_{(0,1/n)}$. Show that (f_n) converges pointwise to zero, but that $\int_{\mathbb{R}} f_n\,d\lambda = 1$ for all $n \in \mathbb{N}$.

Does either of the monotone or dominated convergence theorems apply to this situation?

4.8 Show that if $f \in \mathcal{L}^1_{\mathbb{R}}$ then so is the mapping $x \to \cos(\alpha x)f(x)$, where $\alpha \in \mathbb{R}$. Prove that

$$\lim_{n\to\infty} \int_{\mathbb{R}} \cos(x/n)f(x)dx = \int_{\mathbb{R}} f(x)dx.$$

4.9 Let (S, Σ, m) be a measure space and (A_n) be a sequence of disjoint sets with $A_n \in \Sigma$ for each $n \in \mathbb{N}$. Set $A = \bigcup_{n=1}^{\infty} A_n$ and let $f : S \to \mathbb{R}$ be measurable. Show that

$$\int_A |f|\,dm = \sum_{n=1}^{\infty} \int_{A_n} |f|\,dm.$$

Hint: Use the monotone convergence theorem.

4.10 Let $(a_n)_{n\in\mathbb{N}}$ be a real valued sequence, viewed as a function $a : \mathbb{N} \to \mathbb{R}$ with $a_n = a(n)$. We work over the measure space $(\mathbb{N}, \mathcal{P}(\mathbb{N}), \#)$, where $\#$ denotes counting measure.

(a) Suppose that $a_n \geq 0$ and fix $N \in \mathbb{N}$. Let $a^{(N)}_n = \mathbb{1}_{\{n\leq N\}}a_n$. Show that $a^{(N)}$ is a simple function, write down its integral, and use the monotone convergence theorem to deduce that

$$\int_{\mathbb{N}} a\,d\# = \sum_{n=1}^{\infty} a_n. \tag{4.25}$$

(b) Now consider a general $a = (a_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}$. Explain briefly why $a \in \mathcal{L}^1(\mathbb{N})$ if and only if $\sum_n |a_n| < \infty$ and deduce that (4.25) holds in this case too.

4.11 Show that $f \overset{\text{a.e.}}{=} g$ defines an equivalence relation on the set of all real-valued measurable functions defined on (S, Σ, m).

4.12 $(\star)$ Prove the *reverse Fatou lemma*: if (f_n) is a sequence of non-negative measurable functions for which $f_n \leq f$ for all $n \in \mathbb{N}$, where $f \in \mathcal{L}^1$, then

$$\limsup_{n\to\infty} \int_S f_n\,dm \leq \int_S \limsup_{n\to\infty} f_n\,dm.$$

Hint: Apply Lemma 4.6.3 to $f - f_n$.

On integration of complex valued functions

4.13 (Δ) Write down a version of the dominated convergence theorem applicable to functions $f : S \to \mathbb{C}$. Prove it using the real case.

4.14 (Δ) Let $a \in \mathbb{R}$. Calculate the value of $\int_0^x e^{iay}\,dy$.

4.15 (Δ) Of Exercises 4.5, 4.9, 4.10 and 4.11, which of these results have natural extensions to complex valued functions? Justify your answers briefly.

You will need to solve those exercises first!

Challenge questions

4.16 Let (S, Σ, m) be a measure space and $f : [a, b] \times S \to \mathbb{R}$ be a measurable function for which

(i) The mapping $x \to f(t, x)$ is in $\mathcal{L}^1$ for all $t \in [a, b]$,

(ii) The mapping $t \to f(t, x)$ is continuous for all $x \in S$,

(iii) There exists $g \in \mathcal{L}^1$ such that $|f(t, x)| \le g(x)$ for all $t \in [a, b], x \in S$.

Use the dominated convergence theorem to show that the mapping $t \to \int_S f(t, x)\, dm(x)$ is continuous at all $t \in [a, b]$.

Hint: Use continuity in terms of sequences, that is show that $\lim_{n\to\infty} \int_S f(t_n, x)\, dm(x) = \int_S f(t, x)\, dm(x)$ *for any sequence* (t_n) *satisfying* $\lim_{n\to\infty} t_n = t$.

4.17 Let (S, Σ, m) be a measure space and $f : [a, b] \times S \to \mathbb{R}$ be a measurable function for which

(i) The mapping $x \to f(t, x)$ is in $\mathcal{L}^1$ for all $t \in [a, b]$,

(ii) The mapping $t \to f(t, x)$ is differentiable for all $x \in S$,

(iii) There exists $h \in \mathcal{L}^1$ such that $\left|\dfrac{\partial f(t, x)}{\partial t}\right| \le h(x)$ for all $t \in [a, b], x \in S$.

Show that the mapping $t \to \int_S f(t, x)\, dm(x)$ is differentiable on (a, b) and that

$$\frac{\partial}{\partial t} \int_S f(t, x)\, dm(x) = \int_S \frac{\partial f(t, x)}{\partial t}\, dm(x).$$

Hint: Use the mean value theorem.

4.18 Let

$$\begin{aligned} f(x) &= -2xe^{-x^2} \\ f_n(x) &= \sum_{r=1}^{n} \left(-2r^2 x e^{-r^2x^2} + 2(r+1)^2 x e^{-(r+1)^2x^2}\right) \end{aligned}$$

for all $x \in \mathbb{R}$.

(a) Show that $f(x) = \lim_{n\to\infty} f_n(x)$ for all $x \in \mathbb{R}$.

(b) Let $a > 0$. Show that f and f_n are Riemann integrable over $[0, a]$ for all $n \in \mathbb{N}$ but that

$$\int_0^a f(x)\, dx \ne \lim_{n\to\infty} \int_0^a f_n(x)\, dx.$$

Neither the monotone or dominated convergence theorems can be used here (follow up exercise: explain why not). This example illustrates that things can go badly wrong without them, even when $f_n(x) \to f(x)$ *for all* x.

Additional questions ($\star$)

These questions explore the definition and properties of the Fourier transform. They are off syllabus but you may find them interesting. They involve integration in $\mathbb{C}$, as described in Section 4.8, and you will need extensions of several key results (e.g. linearity, dominated convergence) to that setting.

4.19 Let $f : \mathbb{R} \to \mathbb{R}$. If $f \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$, where λ is Lebesgue measure, define its *Fourier transform* $\widehat{f}(y)$ for each $y \in \mathbb{R}$, by

$$\begin{aligned}\widehat{f}(y) &= \int_{\mathbb{R}} e^{-ixy} f(x)dx \\ &= \int_{\mathbb{R}} \cos(xy)f(x)dx - i\int_{\mathbb{R}} \sin(xy)f(x)dx.\end{aligned}$$

Prove that $|\widehat{f}(y)| < \infty$ and so $\widehat{f}$ is a well-defined function from $\mathbb{R}$ to $\mathbb{C}$. Show also that the Fourier transformation $\mathcal{F}f = \widehat{f}$ is linear, i.e. for all $f, g \in \mathcal{L}^1$ and $a, b \in \mathbb{R}$ we have

$$\widehat{af + bg} = a\widehat{f} = b\widehat{g}.$$

4.20 Recall Dirichlet's jump function $\mathbb{1}_{\mathbb{Q}}$. Does it make sense to write down the Fourier coefficients $a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} \mathbb{1}_{\mathbb{Q}}(x)\cos(nx)dx$ and $b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} \mathbb{1}_{\mathbb{Q}}(x)\sin(nx)dx$ as Lebesgue integrals? If so, what values do they have? Can you associate a Fourier series to $\mathbb{1}_{\mathbb{Q}}$? If so, (and if it is convergent) what does it converge to?

4.21 Fix $a \in \mathbb{R}$ and define the shifted function $f_a(x) = f(x - a)$. If $f \in \mathcal{L}^1$, show that $f_a \in \mathcal{L}^1$, and deduce that $\widehat{f_a}(y) = e^{-iay}\widehat{f}(y)$ for all $y \in \mathbb{R}$.

4.22 Show that the mapping $y \to \widehat{f}(y)$ is continuous from $\mathbb{R}$ to $\mathbb{C}$.

4.23 Suppose that the mappings $x \to f(x)$ and $x \to xf(x)$ are both in $\mathcal{L}^1$. Show that $y \to \widehat{f}(y)$ is differentiable and that for all $y \in \mathbb{R}$,

$$(\widehat{f})'(y) = -i\widehat{g}(y),$$

where $g(x) = xf(x)$ for all $x \in \mathbb{R}$.

Hint: Use the inequality $|e^{ib} - 1| \leq |b|$ for $b \in \mathbb{R}$.

4.24 Assume that $f, g \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ and that g is bounded. Define the *convolution* $f * g$ of f with g by

$$(f * g)(x) = \int_{\mathbb{R}} f(x - y)g(y)dy,$$

for all $x \in \mathbb{R}$. Show that $|(f * g)(x)| < \infty$, and so $f * g$ is a well–defined function from $\mathbb{R}$ to $\mathbb{R}$. Show further that $f * g \in \mathcal{L}^1$, and that the Fourier transform of the convolution is the product of the Fourier transforms, i.e. that for all $y \in \mathbb{R}$,

$$\widehat{f * g}(y) = \widehat{f}(y)\widehat{g}(y).$$

Remark 4.11.1 Analogues of the results of Problems 4.19-4.24, with slight modifications, also hold for the *Laplace transform* $\mathcal{L}f(y) = \int_0^\infty e^{-yx}f(x)dx$, where $y \geq 0$ and $x \mapsto e^{-yx}f(x)$ is assumed to be in $\mathcal{L}^1((0, \infty))$.

Chapter 5

Probability with Measure

In this chapter we will examine probability theory from the measure theoretic perspective. The realisation that measure theory is the foundation of probability is due to the Russian mathematician A. N. Kolmogorov (1903-1987) who in 1933 published the hugely influential "Grundbegriffe der Wahrscheinlichkeitsrechnung" (in English: *Foundations of the Theory of Probability*). Since that time, measure theory has underpinned all mathematically rigorous work in probability theory and has been a vital tool in enabling the theory to develop both conceptually and in applications.

We have already noted that a probability is a measure, random variables are measurable functions and expectation is a Lebesgue integral – but it is not fair to claim that "probability theory" can be reduced to a subset of "measure theory". This is because in probability we model chance and unpredictability, which brings in a set of intuitions and ideas that go well beyond those of weights and measures.

The Polish mathematician Mark Kac (1914-1984) famously described probability theory as "measure theory with a soul." A less eloquent observation is that the notation tends to be much easier to handle in probability. We introduced probability measures as an example in Section 1.3.1, but let us give a formal definition here.

Definition 5.0.1 A measure m is said to be a probability measure if it has total mass 1.

A measure space (S, Σ, m) is said to be a *probability space* if m is a probability measure.

5.1 Probability

In Chapters 5-6 we will work over general *probability spaces* of the form $(\Omega, \mathcal{F}, \mathbb{P})$. An *event* is a measurable set $A \in \mathcal{F}$. We have $\mathbb{P}[\Omega] = 1$ so

$$\mathbb{P}[\Omega] = 1 \quad \text{and} \quad 0 \leq \mathbb{P}[A] \leq 1 \text{ for all } A \in \mathcal{F}.$$

Intuitively, $\mathbb{P}[A]$ is the probability that the event $A \in \mathcal{F}$ takes place. We will generally assign a special status to probability measures and expectations by writing their arguments in square brackets e.g. $\mathbb{P}[A]$ instead of $\mathbb{P}(A)$. This just a convention – there is no difference in mathematical meaning.

In probability we often use 'complement' notation, that is $A^c = \Omega \setminus A$. The standard formulae $\mathbb{P}[A^c] = 1 - \mathbb{P}[A]$ and $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cup B]$ are simply restatements of equations (1.3) and (1.4) in probabilistic notation.

A *random variable* X is a measurable function $X : \Omega \to \mathbb{R}$, where we use the measure spaces from $(\Omega, \mathcal{F}, \mathbb{P})$ and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. If $A \in \mathcal{B}(\mathbb{R})$, it is standard to use the notation $\{X \in A\}$ to denote the event $X^{-1}(A) \in \mathcal{F}$. This allows us to think of X as an object that takes a random value, and this random value might (or might not) fall into the set $A \subseteq \mathbb{R}$. We can thus connect our intuition for probability to the formal machinery of measure theory. Note that we require measurability of the function X and the set A to ensure that the probability $\mathbb{P}[X \in A]$ is defined.

The *law* or *distribution* of X is the induced probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ given by $p_X(B) = \mathbb{P}[X^{-1}(B)]$ for $B \in \mathcal{B}(\mathbb{R})$. Thus

$$p_X(B) = \mathbb{P}[X \in B] = \mathbb{P}[X^{-1}(B)] = \mathbb{P}[\{\omega \in \Omega; X(\omega) \in B\}].$$

The *expectation* of X is the Lebesgue integral

$$\mathbb{E}[X] = \int_\Omega X(\omega)\, d\mathbb{P}(\omega).$$

According to Definition 4.5.1 this is possibly undefined, and when it is defined it is an extended real number. The nice case is when $X \in \mathcal{L}^1$, which occurs precisely when $\mathbb{E}[X] \in \mathbb{R}$. When $\mathbb{E}[X]$ is defined we often write $\mu_X = \mathbb{E}(X)$ and call it the *mean* of X. Note that for all $A \in \mathcal{F}$

$$\mathbb{P}[A] = \mathbb{E}[\mathbb{1}_A]$$

because 1_A is a simple function $\mathbb{1}_A(\omega)$. In probability we often write $\mathbb{1}\{\omega \in A\}$, which we call the indicator of the event A.

By Theorem 3.1.5, essentially anything we can think of doing with random variables will just give us back more random variables. In particular, any Borel measurable function f from $\mathbb{R}$ to $\mathbb{R}$ enables us to construct a new random variable $f(X)$, which is defined pointwise via $f(X)(\omega) = f(X(\omega))$ for all $\omega \in \Omega$. For example we may take $f(x) = x^n$ for any $n \in \mathbb{N}$. This particular choice of f generates the n^{th} moment $\mathbb{E}[X^n]$, which will be a real number if and only if $|X|^n \in \mathcal{L}^1$.

If X has a finite second moment then its *variance* $\text{var}(X) = \mathbb{E}[(X - \mu)^2]$ always exists (see Problem **5.11**). It is common to use the notation $\sigma_X^2 = \text{var}(X)$. The *standard deviation* of X is $\sigma_X = \sqrt{\text{var}(X)}$. When it is clear which random variable we mean, we might write simply μ and σ in place of μ_X, σ_X.

If X and Y are random variables defined on the same probability space and $A, B \in \mathcal{B}(\mathbb{R})$ it is standard to write:

$$\mathbb{P}[X \in A, Y \in B] = \mathbb{P}[\{X \in A\} \cap \{Y \in B\}].$$

For unions we tend to simply write $\mathbb{P}[X \in A \text{ or } Y \in B] = \mathbb{P}[\{X \in A\} \cup \{Y \in B\}]$.

Let us now update the results of Section 1.7 into the language of probability. Recall that a sequence of sets (A_n) with $A_n \in \mathcal{F}$ for all $n \in \mathbb{N}$ is increasing if $A_n \subseteq A_{n+1}$ for all $n \in \mathbb{N}$, and *decreasing* if $A_n \supseteq A_{n+1}$ for all $n \in \mathbb{N}$.

Lemma 5.1.1 *Let $A_n, B_n \in \mathcal{F}$.*

1. *Suppose (A_n) is increasing and $A = \bigcup_n A_n$. Then $\mathbb{P}[A] = \lim_{n\to\infty} \mathbb{P}[A_n]$.*
2. *Suppose (B_n) is decreasing and $B = \bigcap_n B_n$. Then $\mathbb{P}[B] = \lim_{n\to\infty} \mathbb{P}[B_n]$.*

PROOF: This is just Lemma 1.7.1 rewritten in the notation of probability. Note that the condition of part 2 holds automatically here, because in probability all events (i.e. measurable sets) have finite measure. ■

The intuition for the above theorem should be clear. The set A_n gets bigger as $n \to \infty$ and, in doing so, gets ever closer to A; the same is true of their probabilities. Similarly for B_n, which gets smaller and closer to B. This result is a probabilistic analogue of the well known fact that monotone increasing (resp. decreasing) sequences of real numbers converge to the respective sups and infs.

We will study convergence of random variables in Section 7.2. For now, note that in probability we use the term *almost surely* in place of the measure theoretic *almost everywhere*. The meaning is the same, for example $X \overset{\text{a.s.}}{=} Y$ means that $\mathbb{P}[X = Y] = 1$, and $X_n \overset{\text{a.s.}}{\to} X$ means that $\mathbb{P}[X_n \to X] = 1$.

The monotone and dominated convergence theorems, Markov's inequality, all the properties of integrals, and so on, can all be re-written in the language of probability. This is for you to do, with several examples in Exercise **5.1**.

(Δ) Those of you taking MAS61022 can now begin your independent reading of Chapter 6, after solving Exercises **5.1** and **5.2**. Chapter 6 does not depend on the rest of Chapter 5.

5.2 The cumulative distribution function

Let $X : \Omega \to \mathbb{R}$ be a random variable. Its *cumulative distribution function* or *cdf* is the mapping $F_X : \mathbb{R} \to [0,1]$ defined for each $x \in \mathbb{R}$ by

$$F_X(x) = \mathbb{P}[X \leq x].$$

When X is clear from the context we might write F instead of F_X. Note that if $x \leq y$ then $\{X \leq x\} \subseteq \{X \leq y\}$, which by monotonicity of measures implies that $F_X(x) \leq F_X(y)$. That is, the function F_X is monotone increasing. It is straightforward to check that for $x \leq y$ we have

$$\mathbb{P}[X > x] = 1 - F_X(x) \tag{5.1}$$

$$\mathbb{P}[x < X \leq y] = F_X(y) - F_X(x) \tag{5.2}$$

and this is left for you in Exercise 5.4.

Our next result gathers together some analytic properties of F_X. Recall that if $f : \mathbb{R} \to \mathbb{R}$, the *left limit* at x is $\lim_{y\uparrow x} f(y)$, and the *right limit* at x is $\lim_{y\downarrow x} f(y)$. In general left and right limits might not exist, but they always do if the function f is monotonic increasing (or decreasing).

Lemma 5.2.1 *Let X be a random variable having cdf F.*

1. $\mathbb{P}[X = x] = F(x) - \lim_{y\uparrow x} F(y)$
2. *The map $x \to F(x)$ is right continuous: $F(x) = \lim_{y\downarrow x} F(y)$ for all $x \in \mathbb{R}$.*
3. $\lim\limits_{x\to-\infty} F(x) = 0$ *and* $\lim\limits_{x\to\infty} F(x) = 1$.

Proof: We prove the three parts in turn. For the first part, let $x \in \mathbb{R}$ and let (a_n) be a sequence of positive numbers that decreases to zero. For each $n \in \mathbb{N}$ define $B_n = \{x - a_n < X \leq x\}$. Then $B_n \in \mathcal{F}$ and $B_{n+1} \subseteq B_n$, with $\bigcap_n B_n = \{X = x\}$. By Lemma 5.1.1 and (5.2) we have

$$\mathbb{P}[X = x] = \mathbb{P}\left[\bigcap_n B_n\right] = \lim_{n\to\infty} \mathbb{P}[B_n] = F(x) - \lim_{n\to\infty} F(x - a_n),$$

and the result follows.

For the second part, again let $x \in \mathbb{R}$ and let (a_n) be a sequence of positive numbers that decreases to zero. For each $n \in \mathbb{N}$ define $A_n = \{X > x + a_n\}$. Then $B_n \in \mathcal{F}$ and $A_n \subseteq A_{n+1}$, with $\bigcup_n A_n = \{X > x\}$. By Lemma 5.1.1 and (5.1) we have

$$1 - F(x) = \mathbb{P}\left[\bigcup_n A_n\right] = \lim_{n\to\infty} \mathbb{P}[A_n] = 1 - \lim_{n\to\infty} F(x + a_n),$$

and the result follows. ■

Remark 5.2.2 By combining the first two parts of Lemma 5.2.1, we have that $\mathbb{P}[X = x_0] = 0$ if and only if $F_X(x)$ is continuous at $x = x_0$.

Remark 5.2.3 ($\star$) It can be shown that a function $F : \mathbb{R} \to \mathbb{R}$ is the cdf of some random variable X if and only if it is monotone increasing and satisfies properties 2 and 3 of Lemma 5.2.1.

5.3 Discrete and continuous random variables

You will probably recall that many useful random variables are found in two special cases. Formally, we say that a random variable X is a:

1. *continuous random variable* if its cdf F_X is continuous at every point $x \in \mathbb{R}$;
2. *discrete random variable* if F_X has jump discontinuities at a countable set of points and is constant between these jumps.

Note that if F_X is continuous at x then $\mathbb{P}[X = x] = 0$ by Remark 5.2.2. In particular, this applies to all $x \in \mathbb{R}$ for continuous random variables.

Many random random variables are neither discrete nor continuous. They occur in rather mundane ways. For example, suppose that we toss a fair coin, on heads we set $X = 1$ and on tails we set $X = U$ where U is a uniform random variable on $[0, 1]$. The c.d.f. of X is then

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{x}{2} & \text{for } x \in [0, 1) \\ 1 & \text{for } x > 1. \end{cases}$$

Thus X is neither discrete nor continuous. Random variables of this nature after often said to have a *mixed* type, particularly in statistics, because we have used a combination of discrete random variables (the coin toss) and continuous random variables to construct them.

We now point out a technicality that is often forgotten in less rigorous courses: a continuous random variable does not need to have a probability density function! Strictly speaking, those that do have a special name; we say X is an

3. *absolutely continuous random variable* if there exists a measurable function $f_X : \mathbb{R} \to [0, \infty)$ such that $F_X(x) = \int_{-\infty}^{x} f_X(y)dy$ for all $x \in \mathbb{R}$.

The function f_X is called the *probability density function* or *p.d.f.* of X. Since $\mathbb{P}[X \in \mathbb{R}] = 1$ we have $\int_{-\infty}^{\infty} f_X(y)dy = 1$.

Lemma 5.3.1 *Every absolutely continuous random variable is a continuous random variable.*

PROOF: Note that

$$\int_{-\infty}^{x} f_X(y)\,dy = \int_{\mathbb{R}} \mathbb{1}_{(y \le x)} f_X(y)\,dy.$$

We want to prove this is a continuous function of x, for which we'll use the dominated convergence theorem and the definition of continuity in terms of sequences. Let $(x_n) \subseteq \mathbb{R}$ be any sequence such that $x_n \to x$. Note that $|\mathbb{1}_{(y \le x_n)} f_X(y)| \le |f_X(y)|$ with $\int_{\mathbb{R}} |f_X(y)|\,dy = \mathbb{P}[X \in \mathbb{R}] = 1 < \infty$. If $y \neq x$ then we have $\mathbb{1}_{(y \le x_n)} \to \mathbb{1}_{(y \le x)}$ which means that $\mathbb{1}_{(y \le x_n)} f_X(y) \to \mathbb{1}_{(y \le x)} f_X(y)$ for almost all $y \in \mathbb{R}$. Hence by the dominated convergence theorem we have $\int_{-\infty}^{x_n} f_X(y)\,dy \to \int_{-\infty}^{x} f_X(y)\,dy$ as $n \to \infty$, as required. ■

It is rare to come across examples of continuous random variables that are not absolutely continuous, but they do exist. For practical purposes most useful continuous random variables are absolutely continuous; examples that you may have encountered previously include the uniform, exponential, normal, Student t, gamma and beta distributions. Typical examples of discrete random variables are the binomial, geometric and Poisson distributions.

5.4 Independence

In this subsection we consider the meaning of independence for *infinite sequences* of events and random variables. A useful heuristic is 'independence means multiply'. Recall that two events $A_1, A_2 \in \mathcal{F}$ are *independent* if

$$\mathbb{P}[A_1 \cap A_2] = \mathbb{P}[A_1]\mathbb{P}[A_2].$$

For three events we would use $\mathbb{P}[A_1 \cap A_2 \cap A_3] = \mathbb{P}[A_1]\mathbb{P}[A_2]\mathbb{P}[A_3]$ and so on.

For many applications, we want to discuss independence of infinitely many events, or to be precise a sequence (A_n) of events with $A_n \in \mathcal{F}$ for all $n \in \mathbb{N}$. The definition of independence is extended from the finite case by considering all finite subsets of the sequence. Formally:

Definition 5.4.1 We say that the events in the sequence (A_n) are *independent* if the finite set $\{A_{i_1}, A_{i_2}, \ldots, A_{i_m}\}$ is independent for all finite subsets $\{i_1, i_2, \ldots, i_m\}$ of the natural numbers, i.e.

$$\mathbb{P}[A_{i_1} \cap A_{i_2} \cap \cdots, A_{i_m}] = \mathbb{P}[A_{i_1}]\mathbb{P}[A_{i_2}]\cdots\mathbb{P}[A_{i_m}].$$

Two random variables X and Y are said to be independent if $\mathbb{P}[X \in A, Y \in B] = \mathbb{P}[X \in A]\mathbb{P}[Y \in B]$ for all $A, B \in \mathcal{B}(\mathbb{R})$. This idea is extended to three or more random variables in the same way as above. For an infinite sequence of random variables (X_n), we say that the X_n are independent if every finite subset $X_{i_1}, X_{i_2}, \ldots, X_{i_m}$ of random variables is independent, i.e.

$$\mathbb{P}[X_{i_1} \in A_{i_1}, X_{i_2} \in A_{i_2}, \ldots, X_{i_m} \in A_{i_m}] = \mathbb{P}[X_{i_1} \in A_{i_1}]\mathbb{P}[X_{i_2} \in A_{i_2}]\cdots\mathbb{P}[X_{i_m} \in A_{i_m}]$$

for all $A_{i_1}, A_{i_2}, \ldots, A_{i_m} \in \mathcal{B}(\mathbb{R})$ and for all finite $\{i_1, i_2, \ldots, i_m\} \subseteq \mathbb{N}$.

We often want to consider random variables in $\mathbb{R}^d$, where $d \in \mathbb{N}$. Let us consider the case $d = 2$. A random variable in $\mathbb{R}^2$ $Z = (X, Y)$ is a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ where $\mathcal{B}(\mathbb{R}^2)$ is the product σ-field introduced in Section 1.9. The law of Z is the function $p_Z(A) = \mathbb{P}[Z \in A]$ where $A \in \mathcal{B}(\mathbb{R}^2)$. The *joint law* of X and Y is $p_Z(A \times B) = \mathbb{P}[X \in A, Y \in B]$ for $A, B \in \mathcal{B}(\mathbb{R})$, and the *marginal laws* of X and Y are $p_X(A) = \mathbb{P}[X \in A]$ and $p_Y(B) = \mathbb{P}[Y \in B]$. From the definitions above, we have that X and Y are independent if and only if

$$p_Z(A \times B) = p_X(A)p_Y(B),$$

i.e. if the joint law factorises as the product of the two marginals. The same ideas extend to $\mathbb{R}^3$ with e.g. $W = (X, Y, Z)$ and so on.

Theorem 5.4.2 *Let X and Y be random variables.*

1. *If X and Y are independent and $f, g : \mathbb{R} \to \mathbb{R}$ are measurable functions then $f(X)$ and $g(Y)$ are independent.*
2. *If $X, Y \in \mathcal{L}^1$ with $XY \in \mathcal{L}^1$ then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*
3. *The following two conditions are equivalent:*
   (a) X and Y are independent;
   (b) $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$ for all bounded measurable functions $f, g : \mathbb{R} \to \mathbb{R}$.

Proof: The first part is left for you in Exercise **5.7**.

($\star$) The proof of the second and third parts uses the result of a tricky exercise and also Fubini's theorem, so we will view them as off-syllabus but we will cover them within lectures. For the second part,

$$\mathbb{E}[XY] = \int_{\mathbb{R}^2} xy\, p_Z(dx, dy) = \left(\int_{\mathbb{R}} x\, p_X(dx)\right)\left(\int_{\mathbb{R}} y\, p_Y(dy)\right) = \mathbb{E}[X]\mathbb{E}[Y]$$

Here, the first equality is the two-dimensional version of Problem **5.13**, and we have used Fubini's theorem (from Section 4.9.1) in the second equality to write the integral over $\mathbb{R}^2$ as a repeated integral.

For the final part, recall that bounded random variables are in $\mathcal{L}^1$, so combining parts 1 and 2 gives that (a)⇒(b). To see that (b)⇒(a), take measurable sets $A, B \in \mathcal{B}(\mathbb{R})$ and set $f = \mathbb{1}_A$ and $g = \mathbb{1}_B$. Then we have $\mathbb{E}[f(X)g(Y)] = \mathbb{P}[X \in A, Y \in B]$ and $\mathbb{E}[f(X)] = \mathbb{P}[X \in A]$, $\mathbb{E}[g(Y)] = \mathbb{P}[Y \in B]$, so (b) gives $\mathbb{P}[X \in A, Y \in B] = \mathbb{P}[X \in A]\mathbb{P}[Y \in B]$. ■

Regarding part 2 of Theorem 5.4.2, note that *dependent* random variables X and Y can also satisfy $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. See Exercise **5.8** for an example of this.

5.5 Exercises on Chapter 5

On probability as measure

5.1 Write down probabilistic versions of the following results, using the notation of probability theory that was introduced in Section 5.1. You should use probability in place of measure, random variables in place of measurable functions, expectation in place of integration, etc.

(a) The monotone and dominated convergence theorems (Theorems 4.3.1 and 4.6.2).

(b) Markov's and Chebyshev's inequalities (Lemma 4.2.3 and Exercise **4.3**).

(c) Theorem 4.4.1.

(d) ($\star$) Fatou's lemma (Lemma 4.6.3).

5.2 Using the version of Chebyshev's inequality that you found in Exercise **5.1**, show that if X is a random variable satisfying $\mathrm{var}(X) < \infty$ then

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq c] \leq \frac{\mathrm{var}(X)}{c^2}.$$

Within probability, this is the most common form in which to apply Chebyshev's inequality.

5.3 Let $a, b \in \mathbb{R}$ with $a < b$ and let U be a continuous uniform random variable on $[a, b]$, which means that the p.d.f. of U is the function $f(u) = \mathbb{1}_{(a,b)}(x)\frac{1}{b-a}$. Let $A \in \mathcal{B}([a, b])$. Find $\mathbb{P}[U \in A]$ in terms of the Lebesgue measure of A.

5.4 Let $X : \Omega \to \mathbb{R}$ be a random variable with cumulative distribution function F.

(a) Deduce that $\mathbb{P}[X > x] = 1 - F(x)$ and $\mathbb{P}[x < X \leq y] = F(y) - F(x)$ for all $x < y$.

(b) Prove the last part of Lemma 5.2.1: show that $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$.

Hint: Use the same method as for the first two parts of the theorem.

5.5 Let X be a random variable. Show that there are at most countably many $x \in \mathbb{R}$ such that $\mathbb{P}[X = x] > 0$.

Hint: What happens to $F_X(x)$ at x such that $\mathbb{P}[X = x] > 0$?

On independence

5.6 (a) Let (A_n) be a sequence of independent events. Show that

$$\mathbb{P}\left[\bigcap_{n\in\mathbb{N}} A_n\right] = \prod_{n=1}^{\infty} \mathbb{P}[A_n]. \tag{5.3}$$

(b) Recall that we define independence of a sequence of events (A_n) in terms of *finite* subsequences (e.g. as in Section 5.4). An 'obvious' alternative definition might to be use (5.3) instead. Why is this not a sensible idea?

5.7 (a) Let A and B be independent events. Show that their complements A^c and B^c are also independent.

(b) Let X and Y be independent random variables and $f,g:\mathbb{R}\to\mathbb{R}$ be Borel measurable. Deduce that $f(X)$ and $g(Y)$ are also independent.

5.8 (a) Let U be a random variable such that $\mathbb{P}[U=-1]=\mathbb{P}[U=1]=\frac{1}{2}$ and let V be a random variable such that $\mathbb{P}[V=0]=\mathbb{P}[V=1]=\frac{1}{2}$, independent of U. Let $X=UV$ and $Y=U(1-V)$. Show that $\mathbb{E}[XY]=\mathbb{E}[X]\mathbb{E}[Y]$ but that X and Y are not independent.

(b) Let X,Y,Z be random variables, where X and Y are independent of each other with $\mathbb{P}[X=1]=\mathbb{P}[X=-1]=\mathbb{P}[Y=1]=-\mathbb{P}[Y=-1]=\frac{1}{2}$, and $Z=XY$. Show that any pair within $\{X,Y,Z\}$ are independent of each other, but that $\{X,Y,Z\}$ is not a set of independent random variables.

On properties of random variables

5.9 Let $M\in[0,\infty)$. Suppose that (X_n) is a sequence of random variables such that for each n we have $|X_n|\le M$, and suppose that $X_n\overset{\text{a.s.}}{\to}X$. Show that $\mathbb{E}[X_n]\to\mathbb{E}[X]$.

5.10 (a) Let X be a random variable that takes values in $\mathbb{N}\cup\{0\}$. Explain why $X=\sum_{i=1}^\infty \mathbb{1}_{\{X\ge i\}}$ and hence show that

$$\mathbb{E}[X]=\sum_{i=1}^\infty \mathbb{P}[X\ge i].$$

(b) Let Y be a random variable taking values in $[0,\infty)$. Use part (a) to deduce that $\sum_{k=1}^\infty \mathbb{P}[Y\ge k]\le\mathbb{E}[Y]\le 1+\sum_{k=1}^\infty \mathbb{P}[Y\ge k]$.

5.11 (a) Suppose that X and Y are random variables and both X^2 and Y^2 are in L^1. Prove the *Cauchy-Schwarz inequality*:

$$|\mathbb{E}[XY]|\le\left(\mathbb{E}[X^2]^{\frac{1}{2}}\right)\left(\mathbb{E}[Y^2]^{\frac{1}{2}}\right).$$

Hint: Consider $g(t)=\mathbb{E}[(X+tY)^2]$ as a quadratic function of $t\in\mathbb{R}$. Note that a quadratic function ax^2+bx+c with at most one real root must satisfy $b^2-4ac\le 0$.

(b) Deduce that if $X^2\in L^1$ then also $X\in L^1$, and in fact $|\mathbb{E}[X]|^2\le\mathbb{E}[X^2]$.

(c) Let X be *any* random variable with a finite mean $\mathbb{E}[X]=\mu$. Show that $\mathbb{E}[X^2]<\infty$ if and only if $\text{var}(X)<\infty$.

5.12 A random variable is said to have an a^{th} *exponential moment* if $\mathbb{E}[e^{a|X|}]<\infty$, where $a>0$.

(a) Let X be a non-negative random variable and $a>0$. Show that $\mathbb{E}[e^{-aX}]\le 1$.

(b) Let X be a random variable with an exponential moment. Show that $\mathbb{E}[|X|^n]<\infty$ for all $n\in\mathbb{N}$.

5.13 Let X be a real-valued random variable with law p_X defined on a probability space $(\Omega,\mathcal{F},\mathbb{P})$. Show that for all bounded measurable functions $f:\mathbb{R}\to\mathbb{R}$,

$$\int_\Omega f(X(\omega))\,d\mathbb{P}(\omega)=\int_{\mathbb{R}} f(x)\,dp_X(x).$$

What can you say about these integrals when f is non-negative but not necessarily bounded?

Hint: Begin with f an indicator function, then extend to simple, bounded non-negative and general bounded measurable functions.

Challenge questions

5.14 (a) Let $\epsilon > 0$. Let (E_n) be a sequence of independent events such that $\mathbb{P}[E_n] \geq \epsilon$ for all $n \in \mathbb{N}$. Show that $\mathbb{P}[\cup_{n\in\mathbb{N}} E_n] = 1$.

(b) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\epsilon > 0$. Suppose that $(E_n)_{n\in\mathbb{N}}$ is a sequence of independent events, with $\mathbb{P}[E_n] \in (\epsilon, 1-\epsilon)$ for all $n \in \mathbb{N}$. Show that $\mathbb{P}[\omega] = 0$ for all $\omega \in \Omega$ and, hence, deduce that Ω is uncountable.

Chapter 6

Inequalities for Random Variables (Δ)

We've seen several examples of useful inequalities in Chapters 4 and 5, including Markov's and Chebyshev's inequalities (Lemma 4.2.3 and Exercise **5.2**), Fatou's lemma (Lemma 4.6.3) and the Cauchy-Schwarz inequality (Exercise **5.11**). These inequalities can all be written in terms of general integrals, as well as in terms of random variables and expectations. In this chapter we will include some further examples of inequalities that are useful in probability theory.

We will look only at inequalities that apply to general random variables. It is worth noting that beyond what is included here there are several more specialized types of inequality (such as those that apply only to martingales, or to Markov chains, or to some particular distributions) that are also well known in probability theory.

6.1 Chernoff bounds (Δ)

Chernoff bounds are based on applying Markov's inequality to the exponential function e^{tX}, where X is a random variable, then choosing $t \in \mathbb{R}$ to make the resulting bound be as strong as possible. Perhaps surprisingly, it is known that that this method gives near optimal bounds in many situations.

Lemma 6.1.1 (Generic Chernoff Bound) *Let X be a random variable. Then*

1. *For all $t > 0$ we have $\mathbb{P}[X \geq c] \leq e^{-tc}\mathbb{E}[e^{tX}]$.*
2. *For all $t < 0$ we have $\mathbb{P}[X \leq c] \leq e^{-tc}\mathbb{E}[e^{tX}]$.*

Proof: Note that $e^{tX} \geq 0$ so $\mathbb{E}[e^{tX}]$ is well defined as a non-negative extended real number. We prove the two claims in turn. For the first, for $t > 0$, from Markov's inequality (Lemma 4.2.3) we have

$$\mathbb{P}[X \geq c] = \mathbb{P}[e^{tX} \geq e^{tc}] \leq \frac{1}{e^{tc}}\mathbb{E}[e^{tX}]$$

as required. For the second, note that for $t < 0$ we have $\mathbb{P}[X \leq c] = \mathbb{P}[e^{tX} \geq e^{tc}]$ (using $t < 0$ reverses the inequality) and then proceed as before. ■

The function $t \mapsto \mathbb{E}[e^{tX}]$ is known as the *moment generating function* or m.g.f. of the random variable X. Its value is an extended real number and it is possible that $\mathbb{E}[e^{tX}] = \infty$ for all $t \neq 0$. However, in many cases the moment generating function is finite for t in (at least) some interval $(-\epsilon, \epsilon)$ where $\epsilon > 0$. That is often enough to derive a Chernoff bound.

Example 6.1.2 Let X have a Binomial(n, p) distribution. We will take $p = \frac{1}{2}$, so $\mathbb{E}[X] = \frac{n}{2}$, $\mathrm{var}(X) = \frac{n}{4}$, and $\mathbb{E}[e^{tX}] = (\frac{1}{2} + \frac{1}{2}e^t)^n$. We will derive a Chernoff bound for $\mathbb{P}[X \geq \frac{3n}{4}]$, but first let us try Markov's and Chebyshev's inequalities. Markov's inequality (Lemma 4.2.3) gives us

$$\mathbb{P}\left[X \geq \frac{3n}{4}\right] \leq \frac{1}{(3n/4)}\frac{n}{2} = \frac{8}{6},$$

which is greater than one and therefore provides no information. Chebyshev's inequality (Exercise 5.2) gives

$$\mathbb{P}\left[X \geq \frac{3n}{4}\right] = \mathbb{P}\left[X - \frac{n}{2} \geq \frac{n}{4}\right] \leq \mathbb{P}\left[|X - \mathbb{E}[X]| \geq \frac{n}{4}\right] \leq \frac{1}{(n/4)^2}\,\mathrm{var}(X) = \frac{4}{n}.$$

This is significantly better and tends to zero as $n \to \infty$. Lastly, Chernoff's bound gives us that

$$\mathbb{P}\left[X \geq \frac{3n}{4}\right] \leq e^{-3nt/4}\left(\frac{1}{2} + \frac{1}{2}e^t\right)^n \tag{6.1}$$

for all $t \geq 0$. Choosing t to minimize the right hand side (which can be done via differentiating, finding turning points, and checking for minima) gives that the maximum occurs when $e^t = 3$. Putting this value for e^t into the above equation and simplifying results in

$$\mathbb{P}\left[X \geq \frac{3n}{4}\right] \leq 3^{-3n/4}(2)^n = \left(\frac{2}{3^{3/4}}\right)^n \approx (0.88)^n.$$

This bound converges to zero much faster than the bound in (6.1).

6.2 The Paley-Zygmund inequality (Δ)

Recall that Markov's and Chebyshev's inequalities controlled how large $\mathbb{P}[X \geq c]$ could become, using the moments of the random variable X. The Paley-Zygmund inequality is similar in style, but it seeks to control how large X can become relative to $\mathbb{E}[X]$.

Lemma 6.2.1 (Paley-Zymund Inequality) *Let X be a non-negative random variable and suppose that $0 < \mathbb{E}[X^2] < \infty$. Then for any $\theta \in [0,1]$,*

$$\mathbb{P}[X > \theta\mathbb{E}[X]] \geq (1-\theta)^2 \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}.$$

Proof: Note that $X = X\mathbb{1}_{\{X \leq \theta\mathbb{E}[X]\}} + X\mathbb{1}_{X > \theta\mathbb{E}[X]}$ and take expectations to obtain

$$\mathbb{E}[X] = \mathbb{E}\left[X\mathbb{1}_{\{X \leq \theta\mathbb{E}[X]\}}\right] + \mathbb{E}\left[X\mathbb{1}_{\{X > \theta\mathbb{E}[X]\}}\right]. \tag{6.2}$$

We will bound the two terms on the right hand side of the above. If $\mathbb{1}_{\{X \leq \theta\mathbb{E}[X]\}}$ is non-zero then $X \leq \theta\mathbb{E}[X]$, so also $X\mathbb{1}_{\{X \leq \theta\mathbb{E}[X]\}} \leq \theta\mathbb{E}[X]$. Hence we have $\mathbb{E}[X\mathbb{1}_{X \leq \theta\mathbb{E}[X]}] \leq \theta\mathbb{E}[X]$. For the second term, we apply the Cauchy-Schwarz inequality (Exercise 5.11) and obtain that

$$\mathbb{E}\left[X\mathbb{1}_{\{X > \theta\mathbb{E}[X]\}}\right] \leq \left(\mathbb{E}[X^2]\mathbb{E}[\mathbb{1}^2_{\{X > \theta\mathbb{E}[X]\}}]\right)^{1/2}.$$

Indicator functions are either zero or one, hence $\mathbb{1}_A^2 = \mathbb{1}_A$, and by (4.2) satisfy $\mathbb{E}[\mathbb{1}_A] = \mathbb{P}[A]$. We thus obtain that above equation is bounded above by $(\mathbb{E}[X^2]\mathbb{P}[X > \theta\mathbb{E}[X]])^{1/2}$. Putting all this into (6.2) we obtain

$$\mathbb{E}[X] \leq \theta\mathbb{E}[X] + \left(\mathbb{E}[X^2]\mathbb{P}[X > \theta\mathbb{E}[X]]\right)^{1/2}$$

which rearranges to the required result. ■

The most common application of the Paley-Zymund inequality is to set $\theta = 0$ and obtain

$$\mathbb{P}[X > 0] \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}. \tag{6.3}$$

In combination with upper bounds on $\mathbb{E}[X]$ and lower bounds on $\mathbb{E}[X^2]$, equation (6.3) is often used to show that X is not identically zero. This technique is particularly useful when the random variable X is known to be a limit of some sequence (X_n), and bounds on $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ can be obtained from corresponding bounds on $\mathbb{E}[X_n]$ and $\mathbb{E}[X_n^2]$ via the monotone and dominated convergence theorems.

More generally (6.3) simply provides a lower bound on the probability that some random variable is non-zero. It is often called the *second moment method.*

Example 6.2.2 Consider a graph with n vertices. For each (unordered) pair of distinct vertices (v_1, v_2), the edge from v_1 to v_2 is present in the graph with probability p, independent of all other pairs of vertices. This graph is known as the Erdős-Renyi graph, often denoted by $G(n,p)$. We are interested in whether $G(n,p)$ contains *isolated* vertices, which are vertices that have no edges connected to them, as $n \to \infty$.

Let E_i be the event that vertex i is isolated, and let $Y_n = \sum_{i=1}^n \mathbb{1}_{E_i}$ be the number of isolated vertices. The probability of each vertex being isolated is $\mathbb{P}[E_i] = (1-p)^{n-1}$, so the expected number of isolated vertices is $\mathbb{E}[Y_n] = \sum_{i=1}^n \mathbb{P}[E_i] = n(1-p)^{n-1}$.

A pair of distinct vertices, that is $i \neq j$, is part of $2n-3$ edges (and not $2n-2$, because they are both part of the edge between them), so $\mathbb{P}[E_i \cap E_j] = (1-p)^{2n-3}$. We can calculate the second moment too, with the help of Exercise **2.5**, as

$$\begin{aligned}\mathbb{E}[Y_n^2] &= \sum_{i,j=1}^{n} \mathbb{P}[E_i \cap E_j] \\ &= \sum_{i=1}^{n} \mathbb{P}[E_i] + \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \mathbb{P}[E_i \cap E_j] \\ &= n(1-p)^{n-1} + n(n-1)(1-p)^{2n-3}\end{aligned}$$

Note that since Y takes integer values, $\mathbb{P}[Y > 0] = \mathbb{P}[Y \geq 1]$. The Paley-Zygmund inequality gives that

$$\mathbb{P}[Y \geq 1] \geq \frac{n^2(1-p)^{2n-2}}{n(1-p)^{n-1} + n(n-1)(1-p)^{2n-3}} = \frac{1}{\frac{1}{n}(1-p)^{-n-3} + \frac{n-1}{n}(1-p)^{-1}}.$$

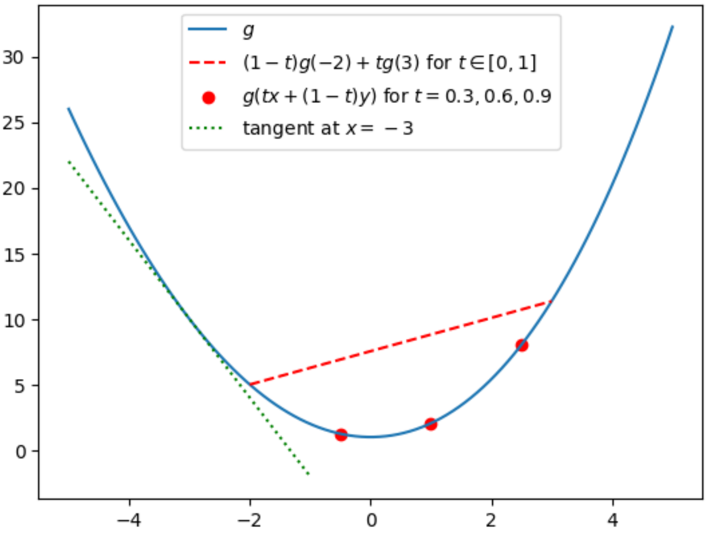
Allowing p to depend on n, it follows that if $\frac{1}{n(1-p)^n} \to 0$, or equivalently $n(1-p)^n \to \infty$, as $n \to \infty$ then we must have $\mathbb{P}[Y \geq 1] \to 1$ as $n \to \infty$. If that condition holds, then for large n it is very likely that $G(n,p)$ contains at least one isolated vertex.

6.3 Jensen's inequality (Δ)

A function $g : \mathbb{R} \to \mathbb{R}$ is *convex* if it satisfies

$$g(tx + (1-t)y) \le tg(x) + (1-t)g(y) \tag{6.4}$$

for all $x, y \in \mathbb{R}$ and $t \in [0, 1]$. Equation (6.4) is best understood with a picture.



Convex functions are often described as having the shape of a smile. The key point is that the red dots, corresponding to the left hand side of (6.4), sit below the red line, corresponding to the right hand side of (6.4). Here we have shown (6.4) for $x = -2$ and $y = 3$, on the convex function $f(x) = 1 + x^2 + \frac{1}{20}\max(0, x^3)$. The tangent line at -3 is shown in green. Note that f sits above its own tangent lines.

Jensen's inequality relates expectation with convex functions. It is a tool that requires a specific situation to apply, but when it does apply it is often the only tool available.

Lemma 6.3.1 (Jensen's Inequality) *Let $X \in \mathcal{L}^1$ and let $g : \mathbb{R} \to \mathbb{R}$ be a convex measurable function. Then $g(\mathbb{E}[X]) \le \mathbb{E}[g(X)]$.*

We will give a proof of Lemma 6.3.1 under the additional assumption that g is differentiable. This is not a very big restriction – in fact, convex functions are necessarily differentiable at all but countably many $x \in \mathbb{R}$ (we won't prove that). However, restricting to differentiable g allows us to give a much more intuitive proof than is possible in the general case. The following lemma explains why, and we'll give the proof of Jensen's inequality below. It says that a differentiable convex function sits above all of its tangent lines, as you can see in the example of the tangent at $x = -3$ pictured in the graph above.

Lemma 6.3.2 *Let $g : \mathbb{R} \to \mathbb{R}$.*

1. *If g is differentiable then g is convex if and only if $g''(x) \ge 0$ for all x.*
2. *If g is convex and differentiable then for all $x, y \in \mathbb{R}$ we have $(y - x)g'(x) \le g(y) - g(x)$.*

PROOF: We'll omit a proof of the first claim. It should help you understand (and check) convexity, but the proof doesn't provide us with any useful intuition. In the picture below (6.4) you can see an example where $g''(x) > 0$ for all x.

For the second claim, we can rewrite (6.4) as $(t + 1 - t)g(tx + (1 - t)y) \le tg(x) + (1 - t)g(y)$, which rearranges to

$$t\left[g\big(tx + (1-t)y\big) - g(x)\right] \le (1-t)\left[g(y) - g\big(tx + (1-t)y\big)\right].$$

Let us first consider when $y > x$. In this case, for $t \neq 1$, dividing through by $(1 - t)(y - x)$ leads to

$$t\frac{g\big(tx + (1-t)y\big) - g(x)}{(1-t)(y-x)} \le \frac{g(y) - g\big(tx + (1-t)y\big)}{(1-t)(y-x)}. \tag{6.5}$$

The point is that $tx + (1 - t)y - x = (1 - t)(y - x)$, so letting $t \uparrow 1$ leads to

$$g'(x) \le \frac{g(y) - g(x)}{y - x}. \tag{6.6}$$

The result follows after multiplying both sides by $y - x$.

When $x < y$ the same calculation leads to the same result, but the direction of the inequality in (6.5) and (6.6) is reversed, then reverses back again when we multiply both sides by $y - x$. The case $x = y$ is trivial. ■

PROOF OF LEMMA 6.3.1 FOR DIFFERENTIABLE g: By part 2 of Lemma 6.3.2 we have $(y - x)g'(x) \le g(y) - g(x)$ for all $x, y \in \mathbb{R}$. Therefore we may put $\mathbb{E}[X]$ in place of x and X in place of y, leading to

$$(X - \mathbb{E}[X])g'(\mathbb{E}[X]) \le g(X) - g(\mathbb{E}[X]).$$

Taking expectations, using linearity and monotonicity,

$$(\mathbb{E}[X] - \mathbb{E}[X])g'(\mathbb{E}[X]) \le \mathbb{E}[g(X)] - g(\mathbb{E}[X]).$$

The left hand side is zero, and the result follows. ■

Example 6.3.3 Setting $g(x) = x^2$ in Jensen's inequality gives $\mathbb{E}[X]^2 \le \mathbb{E}[X^2]$. We derived this inequality already in Exercise **5.11** as a consequence of the Cauchy-Schwarz inequality. Setting $g(x) = |x|^p$ for $p \ge 1$ gives that $\mathbb{E}[|X|^p] \le \mathbb{E}[|X|]^p$, which we will make use of in the proof of Lemma 7.2.1. Note that for odd values of p, the function $x \mapsto |x|^p$ is convex but is not differentiable at $x = 0$.

Remark 6.3.4 Jensen's inequality also holds if the convex function g is only defined on some interval I of $\mathbb{R}$, but I is large enough that $\mathbb{P}[X \in I] = 1$. The proof is the same as above, except that the first inequality in the proof only holds almost surely. Lemma 6.3.2 still holds, but with x and y restricted to I.

Jensen's inequality is surprisingly far reaching in its consequences. For example it provides the key ingredient used to prove most of the basic inequalities of functional analysis e.g. Hölders inequality, Minkowski's inequality, Young's inequality, and so on. The road towards those inequalities is fairly long and begins with Exercise **6.5**, but we won't include those inequalities in this course.

6.4 Exercises on Chapter 6 (Δ)

6.1 Let X_n have a Poisson distribution with mean $n\lambda > 0$. Show that $\mathbb{E}[e^{tX_n}] = e^{n\lambda(e^t-1)}$. Hence construct a Chernoff bound for $\mathbb{P}[X_n > n\lambda^2]$, where $n \in \mathbb{N}$.

6.2 Let $(E_n)_{n\in\mathbb{N}}$ be a sequence of events and let $X_n = \sum_{i=1}^n \mathbb{1}_{E_i}$. Suppose $\sum_{i=1}^\infty \mathbb{P}[E_i] = \infty$ and that $\mathbb{P}[E_i \cap E_j] \leq \mathbb{P}[E_i]\mathbb{P}[E_j]$ for all $i \neq j$. Show that $\mathbb{P}[X_n \geq 1] \to 1$ as $n \to \infty$.

6.3 In each case, determine whether the two quantities given satisfy an equality of the form $a \leq b$, $b \leq a$, or if no such inequality holds in general.

(a) $\mathbb{E}[X^4]$ and $\mathbb{E}[X]^4$, where X is a random variable.

(b) $\mathbb{E}[X^{1/4}]$ and $\mathbb{E}[X]^{1/4}$, where X is a non-negative random variable.

(c) $\mathbb{E}[e^X]$ and $e^{\mathbb{E}[X]}$, where X is a bounded random variable.

(d) $\mathbb{E}[\cos(X)]$ and $\cos(\mathbb{E}[X])$, where X is a random variable.

6.4 Let $\{x_1, \ldots, x_n\} \subseteq (0,\infty)$. The mean average of these values is $\frac{x_1+\ldots x_n}{n}$, which is more precisely known as the *arithmetic mean*. In some situations it is advantageous to instead use the *geometric* mean, $\sqrt[n]{x_1x_2\ldots x_n}$.

By applying Jensen's inequality to a random variable with the discrete uniform distribution on $\{x_1, \ldots, x_m\}$ and the function $g(x) = -\log x$, deduce that

$$\sqrt[n]{x_1x_2\ldots x_n} \leq \frac{x_1 + \ldots x_n}{n}.$$

This equation is known as the AM-GM inequality.

6.5 Let $1 \leq p \leq q$ and let X be a random variable. Use Jensen's inequality to show that if $\mathbb{E}[|X|^q] < \infty$ then $\mathbb{E}[|X|^p] < \infty$.

Challenge questions

6.6 Let X be a non-negative random variable with $\mathbb{E}[X^2] \in (0,\infty)$. Show that $\mathbb{E}[X] > 0$ and that

$$\mathbb{P}[X = 0] \leq \min\left\{\frac{\mathbb{E}[X^2]}{\mathbb{E}[X]^2} - 1,\ 1 - \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}\right\}.$$

Chapter 7

Sequences of Random Variables

In this section we think about sequences of random variables, and about taking limits of random variables. As with real numbers, sequences and limits are our main tool for justifying the use of approximations. Approximations allow us to better understand complicated models, by giving us a way to replace complicated random objects with simpler ones (whilst still maintaining some degree of accuracy). As such, this theory underpins much of stochastic modelling.

7.1 The Borel-Cantelli lemmas

The Borel-Cantelli lemmas are a tool for understanding the tail behaviour of a sequence (E_n) of events. The key definitions are

$$\begin{aligned} \{E_n \text{ i.o.}\} = \{E_n, \text{ infinitely often}\} &= \bigcap_m \bigcup_{n\geq m} E_n = \{\omega : \omega \in E_n \text{ for infinitely many } n\} \\ \{E_n \text{ e.v.}\} = \{E_n, \text{ eventually}\} &= \bigcup_m \bigcap_{n\geq m} E_n = \{\omega : \omega \in E_n \text{ for all sufficiently large } n\}. \end{aligned}$$

The set $\{E_n \text{ i.o.}\}$ is the event that infinitely many of the individual events E_n occur. The set $\{E_n \text{ e.v.}\}$ is the event that, for some (random) N, all the events E_n for which $n \geq N$ occur.

For example, we might take an infinite sequence of coin tosses and choose E_n to be the event that the n^{th} toss is a head. Then $\{E_n \text{ i.o.}\}$ is the event that infinitely many heads occur, and $\{E_n \text{ e.v.}\}$ is the event that, after some point, all remaining tosses show heads.

Note that by straightforward set algebra,

$$\Omega \setminus \{E_n \text{ i.o.}\} = \{\Omega \setminus E_n \text{ e.v.}\}. \tag{7.1}$$

In our coin tossing example, $\Omega \setminus E_n$ is the event that the n^{th} toss is a tail. So (7.1) says that 'there are not infinitely many heads' if and only if 'eventually, we see only tails'.

The Borel-Cantelli lemmas, respectively, give conditions under which the probability of $\{E_n \text{ i.o.}\}$ is either 0 or 1.

Lemma 7.1.1 (First Borel-Cantelli Lemma) *Let $(E_n)_{n\in\mathbb{N}}$ be a sequence of events and suppose $\sum_{n=1}^\infty \mathbb{P}[E_n] < \infty$. Then $\mathbb{P}[E_n \text{ i.o.}] = 0$.*

PROOF: We have

$$\mathbb{P}\left[\bigcap_N \bigcup_{n\geq N} E_n\right] = \lim_{N\to\infty} \mathbb{P}\left[\bigcup_{n\geq N} E_N\right] \leq \lim_{N\to\infty} \sum_{n=N}^\infty \mathbb{P}[E_n] = 0,$$

Here, the first step follows by applying Lemma 5.1.1 to the decreasing sequence of events (B_N) where $B_N = \bigcup_{n\geq N} E_n$. The second stop follows by Lemma 1.7.2 and the fact that limits preserve weak inequalities. The final step follows because $\sum_{n=1}^\infty \mathbb{P}[E_n] < \infty$. ■

For example, suppose that (X_n) are random variables that take the values 0 and 1, and that $\mathbb{P}[X_n = 1] = \frac{1}{n^2}$ for all n. Then $\sum_n \mathbb{P}[X_n = 1] = \sum_n \frac{1}{n^2} < \infty$ so, by Lemma 7.1.1, $\mathbb{P}[X_n = 1 \text{ i.o.}] = 0$, which by (7.1) means that $\mathbb{P}[X_n = 0 \text{ e.v.}] = 1$. So, almost surely, beyond some (randomly • located) point in our sequence (X_n), we will see only zeros. Note that we did not require the (X_n) to be independent.

Lemma 7.1.2 (Second Borel-Cantelli Lemma) *Let $(E_n)_{n\in\mathbb{N}}$ be a sequence of independent events and suppose that $\sum_{n=1}^\infty \mathbb{P}[E_n] = \infty$. Then $\mathbb{P}[E_n \text{ i.o.}] = 1$.*

Proof: Write $E_n^c = \Omega \setminus E_n$. We will show that $\mathbb{P}[E_n^c \text{ e.v.}] = 0$, which by (7.1) implies our stated result. Note that

$$\mathbb{P}[E_n^c \text{ e.v.}] = \mathbb{P}\left[\bigcup_N \bigcap_{n\geq N} E_n^c\right] \leq \sum_{N=1}^\infty \mathbb{P}\left[\bigcap_{n\geq N} E_n^c\right] \tag{7.2}$$

by Lemma 1.7.2. Moreover, since the (E_n) are independent, so are the (E_n^c), so

$$\mathbb{P}\left[\bigcap_{n\geq N} E_n^c\right] = \prod_{n=N}^\infty \mathbb{P}[E_n^c] = \prod_{n=N}^\infty (1 - \mathbb{P}[E_n]) \leq \prod_{n=N}^\infty e^{-\mathbb{P}[E_n]} = \exp\left(-\sum_{n=N}^\infty \mathbb{P}[E_n]\right) = 0.$$

Here, the first step follows by Exercise 5.6. The second step is immediate and the third step uses that $1 - x \leq e^{-x}$ for $x \in [0, 1]$. The fourth step is immediate and the final step holds because $\sum_n \mathbb{P}[E_n] = \infty$. By (7.2) we thus have $\mathbb{P}[E_n^c \text{ e.v.}] = 0$. ■

For example, suppose that (X_n) are i.i.d. random variables such that $\mathbb{P}[X_n = 1] = \frac{1}{2}$ and $\mathbb{P}[X_n = -1] = \frac{1}{2}$. Then $\sum_n \mathbb{P}[X_n = 1] = \infty$ and, by Lemma 7.1.2, $\mathbb{P}[X_n = 1 \text{ i.o.}] = 1$. By symmetry, we have also $\mathbb{P}[X_n = 0 \text{ i.o.}] = 1$. So, if we look along our sequence, almost surely we will see infinitely many 1s and infinitely many 0s.

Since both the Borel-Cantelli lemmas come down to summing a series, a useful fact to remember from real analysis is that, for $p \in \mathbb{R}$,

$$\sum_{n=1}^\infty n^{-p} < \infty \quad \Leftrightarrow \quad p > 1.$$

Recall that this fact follows from the integral test for convergence of series.

7.2 Convergence of random variables

Let (X_n) be a sequence of random variables, all of which are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. There are various different ways in which we can examine the convergence of this sequence to a random variable X (which is also defined on $(\Omega, \mathcal{F}, \mathbb{P})$. They are called *modes of convergence*.

When we talk about convergence of real numbers $a_n \to a$ we only have one mode of convergence, which we might think of as convergence of the value of a_n to the value of a. Random variables are much more complicated objects; they take many different values with different probabilities. For this reason, there are multiple different modes of convergence of random variables.

We have already mentioned almost sure convergence of random variables in Section 5.1. We recall this and introduce three new modes of convergence here. We say that (X_n) converges to X

- *in distribution* if whenever $\mathbb{P}[X = x] = 0$ we have $\mathbb{P}[X_n \leq x] \to \mathbb{P}[X \leq x]$.
- *in probability* if given any $a > 0$, we have $\mathbb{P}[|X_n - X| \geq a] \to 0$ as $n \to \infty$,
- *almost surely* if $\mathbb{P}[X_n \to X, \text{ as } n \to \infty] = 1$,
- *in $\mathcal{L}^p$* if $\mathbb{E}[|X_n - X|^p] \to 0$ as $n \to \infty$.

When (X_n) converges to X almost surely we sometimes write $X_n \to X$ a.s. as $n \to \infty$. We may also write the type of convergence above the arrow e.g. $X_n \xrightarrow{\mathcal{L}^2} X$ or $X_n \xrightarrow{a.s.} X$.

For $\mathcal{L}^p$ convergence we are usually only interested in the cases $p = 1$ and $p = 2$. The case $p = 2$ is sometimes known as convergence in *mean square*. For reasons that are explained in Section 4.9.2, convergence in $\mathcal{L}^p$ is only defined for $p \in [1, \infty)$.

Happily, there are some relationships between these different modes of convergence.

Lemma 7.2.1 *Let X_n, X be random variables.*

1. *If $X_n \xrightarrow{\mathbb{P}} X$ then $X_n \xrightarrow{\mathrm{d}} X$.*
2. *If $X_n \xrightarrow{\mathrm{a.s.}} X$ then $X_n \xrightarrow{\mathbb{P}} X$.*
3. *If $X_n \xrightarrow{\mathcal{L}^p} X$ then $X_n \xrightarrow{\mathbb{P}} X$.*
4. *Let $1 \leq p < q$. If $X_n \xrightarrow{\mathcal{L}^p} X$ then $X_n \xrightarrow{\mathcal{L}^p} X$.*

No other relationships exist in general, other than those implied by the above results.

PROOF: The last part follows from the counterexamples in Exercises 7.3-7.5. We'll give proofs of parts 1-4 here.

Part 1. Let $x \in \mathbb{R}$ be such that $\mathbb{P}[X = x] = 0$ and let $\epsilon > 0$. We have

$$\begin{aligned}\mathbb{P}[X \leq x] &= \mathbb{P}[X \leq x, |X_n - X| < \epsilon] + \mathbb{P}[X \leq x, |X_n - X| \geq \epsilon] \\ &\leq \mathbb{P}[X_n \leq x + \epsilon] + \mathbb{P}[|X_n - X| \geq \epsilon]. \qquad (7.3)\end{aligned}$$

Putting $x - \epsilon$ in place of x we obtain

$$\mathbb{P}[X \leq x - \epsilon] \leq \mathbb{P}[X_n \leq x] + \mathbb{P}[|X_n - X \geq \epsilon]. \qquad (7.4)$$

Combining (7.3) and (7.4) leads to

$$\mathbb{P}[X \le x - \epsilon] - \mathbb{P}[|X_n - X| \ge \epsilon] \;\le\; \mathbb{P}[X_n \le x] \;\le\; \mathbb{P}[X \le x + \epsilon] + \mathbb{P}[|X_n - X| \ge \epsilon].$$

By Remark 5.2.2 and the fact that $\mathbb{P}[X = x] = 0$ we have that $y \mapsto \mathbb{P}[X \le y]$ is continuous at $y = x$. Hence, letting $\epsilon \to 0$, both $\mathbb{P}[X \le x + \epsilon]$ and $\mathbb{P}[X \le x - \epsilon]$ converge to $\mathbb{P}[X \le x]$. Since limits preserve weak inequalities this gives

$$\mathbb{P}[X \le x] - \mathbb{P}[|X_n - X| \ge \epsilon] \;\le\; \mathbb{P}[X_n \le x] \;\le\; \mathbb{P}[X \le x] + \mathbb{P}[|X_n - X| \ge \epsilon].$$

Letting $n \to \infty$ and using that $X_n \xrightarrow{\mathbb{P}} X$, the sandwich rule gives $\mathbb{P}[X_n \le x] \to \mathbb{P}[X \le x]$.

Part 2. Let $\epsilon > 0$ be arbitrary and let $A_n = \bigcup_{m=n}^{\infty} \{|X_m - X| \ge \epsilon\}$. Then (A_n) is a decreasing sequence of events. Let $A = \bigcap_{n=1}^{\infty} A_n$. If $\omega \in A$ then $X_n(\omega)$ cannot converge to $X(\omega)$ as $n \to \infty$ and so $\mathbb{P}[A] = 0$, because $X_n \to X$ almost surely. By Lemma 5.1.1 part (2), $\lim_{n\to\infty} \mathbb{P}[A_n] = \mathbb{P}[A] = 0$. But then by monotonicity,

$$\mathbb{P}[|X_n - X| \ge \epsilon] \le \mathbb{P}[A_n] \to 0 \text{ as } n \to \infty.$$

Part 4. (Δ) We'll prove part 4 next, because it will be helpful in part 3 below. This part is marked with a (Δ) because we need an inequality that comes from Chapter 6. It is true that

$$\mathbb{E}[|X|]^p \le \mathbb{E}[|X|^p] \tag{7.5}$$

for all $p \ge 1$ and all random variables X. For general p, this is a special case of Jensen's inequality, which is part of the independent reading in Section 6.3, marked with a (Δ). However, the $p = 2$ case is a consequence of the Cauchy-Schwartz inequality and has already appeared in Exercise **5.11**. The $p = 1$ case is simply $|\mathbb{E}[X]| \le \mathbb{E}[|X|]$, which is the absolute value property from Theorem 4.5.3 written in the notation of probability.

Putting $|X_n - X|^p$ into (7.5), and then putting $q/p \ge 1$ in place of p, we obtain that $\mathbb{E}[|X_n - X|^p]^{q/p} \le (\mathbb{E}[|X_n - X|^{p(q/p)}])$. Thus $\mathbb{E}[|X_n - X|^p] \le (\mathbb{E}[|X_n - X|^q])^{p/q}$. The result follows.

Part 3. Thanks to part 4, it suffices to prove that L^1 convergence implies convergence in probability. From Markov's inequality (Lemma 4.2.3) for any $a > 0$ we have

$$\mathbb{P}[|X_n - X| \ge a] \le \frac{\mathbb{E}[|X_n - X|]}{a}.$$

If $X_n \xrightarrow{L^1} X$ then the right hand side tends to zero as $n \to \infty$, hence so does the left. ■

Remark 7.2.2 The following diagram records which modes of convergence imply which other modes of convergence.

$$\begin{tikzcd}
 & \text{a.s.} \arrow[dr, Rightarrow] & & \\
 & & \mathbb{P} \arrow[r, Rightarrow] & d \\
L^q \arrow[r, Rightarrow] & L^p \arrow[ur, Rightarrow] & &
\end{tikzcd}$$

$q > p$

In all other cases (i.e. that are not automatically implied by the above), convergence in one mode does not imply convergence in another.

Recall that for real numbers, if $a_n \to a$ and $a_n \to b$ then $a = b$, which is known as uniqueness of limits. For random variables, the situation is a little more complicated: if $X_n \xrightarrow{\mathbb{P}} X$ and $X_n \xrightarrow{\mathbb{P}} Y$ then $X = Y$ almost surely. By Lemma 7.2.1, this result also applies to $\xrightarrow{L^p}$ and $\xrightarrow{a.s.}$. However, if we have only $X_n \xrightarrow{d} X$ and $X_n \xrightarrow{d} Y$ then we can only conclude that X and Y have the same distribution, that is $\mathbb{P}[X \le x] = \mathbb{P}[Y \le x]$ for all x. Proving these facts is Exercise 7.6.

Establishing convergence in distribution, probability and L^p usually comes down to calculating (or estimating) the important quantities involved: $\mathbb{P}[X \le x]$, $\mathbb{P}[|X_n - X| \le a]$ and $\mathbb{E}[|X_n - X|^p]$, and then thinking about their limits as $n \to \infty$. There are several examples of this type within the exercises. Almost sure convergence is harder to work with, but here we can often use the Borel-Cantelli lemmas.

Example 7.2.3 Let (X_n) be a sequence of i.i.d. random variables, each with the uniform distribution on $[0,1]$. Then $\mathbb{P}[X_n \le \frac{1}{3}] = \mathbb{P}[X_n \ge \frac{2}{3}] = \frac{1}{3}$, so (using independence) by two applications of the second Borel-Cantelli lemma we have $\mathbb{P}[X_n \le \frac{1}{3}\text{ i.o.}] = \mathbb{P}[X_n \ge \frac{2}{3}\text{ i.o.}] = 1$. Hence, with probability 1, the sequence X_n will oscillate infinitely often between $[0, \frac{1}{3}]$ and $[\frac{2}{3}, 1]$, in which case it cannot converge. Thus (X_n) does not converge almost surely (to any limit) in this case.

Alternatively, consider a sequence (Y_n) with $\mathbb{P}[Y_n = 1] = \frac{1}{n^2}$ and $\mathbb{P}[Y_n = 0] = 1 - \frac{1}{n^2}$. Since $\sum \frac{1}{n^2} < \infty$, the first Borel-Cantelli lemma tells us that $\mathbb{P}[Y_n = 1\text{ i.o.}] = 0$. Hence $\mathbb{P}[Y_n = 0\text{ e.v.}] = 1$, which implies that $\mathbb{P}[Y_n \to 0] = 1$, or in other words that $Y_n \xrightarrow{a.s.} 0$.

Lemma 7.2.4 *If $X_n \to X$ in probability as $n \to \infty$ then there is a subsequence of (X_n) that converges to X almost surely.*

PROOF: If (X_n) converges in probability to X, for all $c > 0$, given any $\epsilon > 0$, there exists $N(c) \in \mathbb{N}$ so that for all $n \ge N(c)$, $\mathbb{P}[|X_n - X| > c] < \epsilon$.

In order to find our subsequence:

– First choose, $c = 1$ and $\epsilon = 1/2$, then for $n \ge N(1)$, $\mathbb{P}[|X_n - X| > 1] < 1/2$.

– Next choose $c = 1/2$ and $\epsilon = 1/4$, then for $n \ge N(2)$, $\mathbb{P}[|X_n - X| > 1/2] < 1/4$.

– In general, choose $c = 1/r$ and $\epsilon = 1/2^r$, then for $n \ge N(r)$, $\mathbb{P}[|X_n - X| > 1/r] < 1/2^r$.

Set $k_r = \max\{N(1), N(2), \dots, N(r), r\}$ for $r \in \mathbb{N}$. to obtain a subsequence (X_{k_r}) so that for all $r \in \mathbb{N}$,

$$\mathbb{P}[|X_{k_r} - X| > 1/r] < 1/2^r.$$

Since $\sum \frac{1}{2^r} < \infty$, by the first Borel-Cantelli lemma (Lemma 7.1.1) we have

$$\mathbb{P}[|X_{k_r} - X| > 1/r \;\; i.o.] = 0,$$

and so

$$\mathbb{P}[|X_{k_r} - X| \le 1/r \;\; e.v.] = 1.$$

Hence, almost surely, there exists some $R \in \mathbb{N}$ such that for all $r \ge R$ we have $|X_{k_r} - X| < 1/r$, which implies that $|X_{k_r} - X| \to 0$. Hence $X_{k_r} \to X$ almost surely. ■

Remark 7.2.5 A subtle point is that the subsequence of (X_n) constructed by Lemma 7.2.4 is a deterministic subsequence, in the sense that $r \mapsto k_r$ is a deterministic function. Of course, X_{k_r} is a random variable for each r.

7.3 Laws of large numbers

Let (X_n) be a sequence of random variables all defined on the same probability space. We say that the sequence (X_n) is *i.i.d.*, short for *independent and identically distributed*, if it has the following properties:

- they are independent;
- they are identically distributed, which means that for all $A \in \mathcal{B}(\mathbb{R})$,
$$\mathbb{P}[X_1 \in A] = \mathbb{P}[X_2 \in A] = \cdots = \mathbb{P}[X_n \in A] = \cdots$$
or equivalently that $p_{X_n} = p_{X_m}$ for all $n \neq m$.

Sequences of this type are very important. From a practical point of view they correspond to a random trial repeated multiple times, which is a common occurrence in experimental science, but also in statistics (think of e.g. survey responses). They are also theoretically important, because the large amount of independence involved makes these sequences easier to study. There are many examples of limit theorems in probability that were proved first for i.i.d. sequences and later extended to more complicated situations. We will study two examples of this in Sections 7.3-7.5.

In this section we are interested in the *empirical arithmetic mean*

$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n}, \tag{7.6}$$

which in statistics is often known as the sample mean. The standard notation $\overline{X_n}$ is rather lazy, strictly $\overline{(\,\cdot\,)}$ is a function whose arguments are the sequence $X = (X_n)$ and the natural number n, with value given by (7.6).

If $X_n \in \mathcal{L}^1$ for some (and hence all) $n \in \mathbb{N}$, with $\mu = \mathbb{E}[X_n]$, then by linearity $\overline{X_n} \in \mathcal{L}^1$ and $\mathbb{E}[\overline{X_n}] = \mu$. It is natural to expect that for the large n the value of $\overline{X_n}$ is typically close to μ, because some of the X_i will fall above μ, others below, and to some extent they will balance each other out. The following two famous results make this idea precise.

Theorem 7.3.1 (Weak Law of Large Numbers) *Let (X_n) be a sequence of i.i.d. random variables with $\mathbb{E}[X_n] = \mu$ for all $n \in \mathbb{N}$. Suppose that $\mathbb{E}[X_n^2] < \infty$ for all $n \in \mathbb{N}$. Then $\overline{X_n} \to \mu$ in probability as $n \to \infty$.*

Proof: By the Cauchy-Schwarz inequality (Exercise 5.11) $\mathbb{E}[|X_nX_m|] \leq (\mathbb{E}[X_n^2]\mathbb{E}[X_m^2]^{1/2} < \infty$, so $X_nX_m \in \mathcal{L}^1$. Also by this exercise we have $\mathrm{var}(X_n) < \infty$, so let us write $\sigma^2 = \mathrm{var}(X_n)$.

Since the (X_n) are independent by Theorem 5.4.2 we have $\mathbb{E}[X_nX_m] = \mathbb{E}[X_n]\mathbb{E}[X_m] = 0$, so $\mathrm{cov}(X_n, X_m) = \mathbb{E}[X_nX_m] - \mathbb{E}[X_n]\mathbb{E}[X_m] = 0$. We may therefore calculate

$$\mathrm{var}(\overline{X_n}) = \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathrm{var}(X_i) + \sum_{i\neq j}\mathrm{cov}(X_i, X_j)\right) = \frac{n\sigma^2}{n^2} + 0 = \frac{\sigma^2}{n}.$$

Hence, by Chebychev's inequality (Exercise 5.2) for all $a > 0$ we have that $\mathbb{P}[|\overline{X_n} - \mu| > a] \leq \frac{\mathrm{var}(\overline{X_n})}{a^2} = \frac{\sigma^2}{na^2}$ which tends to zero as $n \to \infty$. ■

Theorem 7.3.2 (Strong Law of Large Numbers) *Let (X_n) be a sequence of i.i.d. random variables $X_n \in \mathcal{L}^1$ and with $\mathbb{E}[X_n] = \mu$ for all $n \in \mathbb{N}$. Then $\overline{X_n} \to \mu$ almost surely as $n \to \infty$.*

By Lemma 7.2.1 the strong law implies the weak law, but the strong law is much harder to prove. We won't give a full proof in this course. Instead, we will give a proof in the special case that $\mathbb{E}[X_n^4] < \infty$.

Proof of Theorem 7.3.2, assuming that $\mathbb{E}[X_n^4] < \infty$: Without loss of generality we may assume that $\mu = 0$. The general case can be obtained from this special case by considering $X_n' = X_n - \mu$.

Let $S_n = X_1 + X_2 + \dots + X_n$ so that $S_n = n\overline{X_n}$ for all $n \in \mathbb{N}$. Consider $\mathbb{E}[S_n^4]$. It contains many terms of the form $\mathbb{E}[X_j X_k X_l X_m]$. By the same argument as we used in the proof of Theorem 7.3.1, based on the Cauchy-Schwarz inequality, $X_j X_k X_l X_m \in \mathcal{L}^1$ for all j, k, l, m. If j, k, l and m are all distinct then by Theorem 5.4.2 we have $\mathbb{E}[X_j X_k X_l X_m] = \mathbb{E}[X_j]\mathbb{E}[X_k]\mathbb{E}[X_l]\mathbb{E}[X_m]$, which is zero because $\mathbb{E}[X_j] = \mathbb{E}[X_k] = \mathbb{E}[X_l] = \mathbb{E}[X_m] = 0$. A similar argument disposes of terms of the form $\mathbb{E}[X_j X_k^3]$ and $\mathbb{E}[X_j X_k X_l^2]$, where j, k, l are distinct.

The only terms with non-vanishing expectation are n terms of the form X_i^4 and $\binom{n}{2}\binom{4}{2} = 3n(n-1)$ terms of the form $X_i^2 X_j^2$ with $i \neq j$. By part 1 of Theorem 5.4.2, X_i^2 and X_j^2 are independent for $i \neq j$ and so by part 2 Theorem 5.4.2

$$\mathbb{E}[X_i^2 X_j^2] = \mathbb{E}[X_i^2]E[X_j^2] = \operatorname{var}(X_i^2)\operatorname{var}(X_j^2) = \sigma^4.$$

Putting all this together and writing $b = \mathbb{E}[X_n^4]$,

$$\begin{aligned}\mathbb{E}[S_n^4] &= \sum_{i=1}^n \mathbb{E}[X_i^4] + \sum_{i \neq j} \mathbb{E}[X_i^2 X_j^2] \\ &= nb + 3n(n-1)\sigma^4 \leq Kn^2,\end{aligned}$$

where $K = nb + 3\sigma^4$. For all $a > 0$, by Markov's inequality (Lemma 4.2.3)

$$\mathbb{P}[|\overline{X_n}| > a] = \mathbb{P}[S_n^4 > a^4 n^4] \leq \frac{\mathbb{E}[S_n^4]}{a^4 n^4} \leq \frac{Kn^2}{a^4 n^4} = \frac{K}{a^4 n^2}.$$

Recall that $\sum_{n=1}^\infty \frac{1}{n^2} < \infty$. Hence, by the first Borel-Cantelli lemma, $\mathbb{P}[|\overline{X_n}| > a \text{ i.o.}] = 0$ and so $\mathbb{P}[|\overline{X_n}| \leq a \text{ e.v.}] = 1$. It follows that $\overline{X_n} \to 0$ a.s. as required. ■

Remark 7.3.3 ($\star$) The proof, in the general case without Assumption 4.1, uses a 'truncation argument' based on $Y_n = X_n \mathbb{1}_{\{X_n \leq n\}}$. Note that $Y_n \leq n$ for all n and so $\mathbb{E}[Y_n^k] \leq n^k$ for all k. If $X_n \geq 0$ for all n, $\mathbb{E}[Y_n] \to \mu$ by monotone convergence. Roughly speaking the argument is to prove a SLLN for the $\overline{Y_n}$, then transfer this to the $\overline{X_n}$s. It requires much more work. You can find it in e.g. **.

7.4 Characteristic functions (Δ)

In this section we introduce the main tool that we will need to prove the central limit theorem. It relies on Lebesgue integration in $\mathbb{C}$, which we studied in Section 4.8. In this section we will also use complex versions of exercises that were earlier set for real valued functions. In such cases the complex version follows by applying the real version to real and imaginary parts.

Definition 7.4.1 Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The *characteristic function* $\phi_X : \mathbb{R} \to \mathbb{C}$ of X and is defined, for each $u \in \mathbb{R}$, by

$$\phi_X(u) = \mathbb{E}\left[e^{iuX}\right] = \int_{\mathbb{R}} e^{iuy}\, dp_X(y). \tag{7.7}$$

The integral formula for $\mathbb{E}[e^{iuX}]$ on the right hand side of (7.7) follows by Exercise **5.13**. Note also that $y \to e^{iuy}$ is measurable since $e^{iuy} = \cos(uy) + i\sin(uy)$, and in $\mathcal{L}^1$ by Exercise **4.5** since $|e^{iuy}| \leq 1$ for all $y \in \mathbb{R}$ and p_X is a finite measure.

Example 7.4.2 Suppose that $X \sim N(\mu, \sigma^2)$, that is X has a normal distribution (sometimes known as a Gaussian) with mean μ and variance σ^2. In Problem **7.9** you can show for yourself that in this case $\phi_X(u) = \exp\left(i\mu u - \frac{1}{2}\sigma^2 u^2\right)$ for all $u \in \mathbb{R}$.

Equation (7.7) states that the characteristic function of X is the Fourier transform of the law p_X of the random variable X. In elementary probability theory courses we often meet the Laplace transform $\mathbb{E}[e^{uX}]$ of X, which is called the *moment generating function* of X. The moment generating function has the disadvantage that it only exists when $|u|$ is small enough, because the function $y \mapsto e^{uy}$ may not be $\mathcal{L}^1$. How small $|u|$ is required to be depends on X, but unfortunately there are random variables for which $\mathbb{E}[e^{uX}]$ is undefined for all $u \neq 0$. In such cases the moment generating function is useless. The characteristic function has the important advantage that it is always defined for all $u \in \mathbb{R}$.

Here is a useful property of characteristic functions, which is another instance of the 'independence means multiply' philosophy that we developed in Section 5.4.

Lemma 7.4.3 *If X and Y are independent random variables then for all $u \in \mathbb{R}$,*

$$\phi_{X+Y}(u) = \phi_X(u)\phi_Y(u).$$

PROOF: We have $\phi_{X+Y}(u) = \mathbb{E}[e^{iu(X+Y)}] = \mathbb{E}\left[e^{iuX}e^{iuY}\right] = \mathbb{E}\left[e^{iuX}\right]\mathbb{E}\left[e^{iuY}\right] = \phi_X(u)\phi_Y(u)$, where the key step follows by the complex version of Theorem 5.4.2. ■

We'll end this section with two important properties of characteristic functions, neither of which will be proved within this course. The first says that characteristics are unique to random variables, in the sense that any random variables with same characteristic function also have the same law. The second relates characteristic functions to convergence in distribution.

Theorem 7.4.4 *If X and Y are two random variables for which $\phi_X(u) = \phi_Y(u)$ for all $u \in \mathbb{R}$ then $p_X = p_Y$.*

Theorem 7.4.5 *Let X_n, X be random variables with laws p_{X_n} and p_X (respectively), and characteristic functions ϕ_n and ϕ. The following statements are equivalent:*

1. *$X_n \xrightarrow{d} X$,*

2. *for all $u \in \mathbb{R}$ we have $\phi_n(u) \to \phi(u)$.*

7.4.1 Approximating characteristic functions with polynomials (Δ)

We now an inequality that we will be used in our proof of the central limit theorem. Let $x \in \mathbb{R}$ and let $R_n(x)$ be the remainder term in the Taylor series expansion of e^{ix},

$$R_n(x) = e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!}.$$

We start with an upper bound on $R_n(x)$, which we then convert into a bound on the distance between a characteristic function and an approximating polynomial.

Lemma 7.4.6 *For all $n \in \mathbb{N} \cup \{0\}$ and $x \in \mathbb{R}$,*

$$|R_n(x)| \leq \min\left\{\frac{2|x|^n}{n!}, \frac{|x|^{n+1}}{(n+1)!}\right\}.$$

Proof: A simple calculation based on Exercise **4.14** shows that

$$R_0(x) = e^{ix} - 1 = \begin{cases} \int_0^x ie^{iy}\,dy & \text{if } x > 0, \\ -\int_x^0 ie^{iy}\,dy & \text{if } x < 0. \end{cases}$$

Using that $|ie^{iy}| \leq 1$ and the absolute values property of integrals (from the complex version of Theorem 4.5.3) gives that $|R_0(x)| \leq |x|$. By the periodicity of sin and the fact that $\int_0^{2\pi} \sin y\,dy = 0$, we have that for all $x \in \mathbb{R}$, $|\int_0^x \sin y\,dy| \leq \int_0^\pi \sin y\,dy = 2$. The same applies to cos. Noting that $e^{iy} = \cos y + i \sin y$, we obtain that $|R_0(x)| \leq (2+2)^{1/2} = 2$. Putting all this together, we have $|R_0(x)| \leq \min\{2, |x|\}$.

A slightly longer calculation, again using on Exercise **4.14**, shows that

$$R_n(x) = \begin{cases} \int_0^x iR_{n-1}(y)\,dy & \text{if } x > 0, \\ -\int_x^0 iR_{n-1}(y)\,dy & \text{if } x < 0. \end{cases}$$

Using this relationship, the result can be shown via induction (which is left for you to check), starting from the base case $n = 0$ above. ∎

Lemma 7.4.7 *Let X be a random variable such that $\mathbb{E}[|X|^n] < \infty$ and $\mathbb{E}[X] = 0$. Let ϕ be the characteristic function of X. Then*

$$\left|\phi(y) - \sum_{k=0}^{n} \frac{(iy)^k \mathbb{E}[X^k]}{k!}\right| \leq \mathbb{E}\left[\min\left\{\frac{2|yX|^n}{n!}, \frac{|yX|^{n+1}}{(n+1)!}\right\}\right].$$

for all $n \in \mathbb{N} \cup \{0\}$.

Proof: We have

$$|\mathbb{E}[R_n(yX)]| \leq \mathbb{E}[|R_n(yX)|] \leq \mathbb{E}\left[\max\left\{\frac{2|yX|^2}{n!}, \frac{|yX|^{n+1}}{(n+1)!}\right\}\right].$$

The first step uses the absolute value property of integrals, and the second step uses Lemma 7.4.6 (with $x = yX$) and monotonicity. The required result follows after noting that by linearity $\mathbb{E}[R_n(yX)] = \phi(y) - \sum_{k=0}^{n} \frac{(iy)^k \mathbb{E}[X^k]}{k!}$. We may use linearity here because, by Exercise **6.5**, if $\mathbb{E}[|X|^n] < \infty$ then $\mathbb{E}[|X|^k] < \infty$ for all $1 \leq k \leq n$. ∎

7.5 The central limit theorem (Δ)

The law of large numbers in Section 7.3 told us that if (X_n) was an i.i.d. sequence of $\mathcal{L}^1$ random variables then the sample mean $\overline{X_n}$ because close to $\mu = \mathbb{E}[X_n]$, for large n. The central limit theorem, which we study in this section, examines how close they become.

Let $S_n = \sum_{i=1}^n X_n$, and let us write

$$Y_n = \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \tag{7.8}$$

where $\mu = \mathbb{E}[X_n]$ and $\sigma = \mathrm{var}(X_n)$, both assumed to be finite. We saw in Section 7.3 that $\mathbb{E}[\overline{X_n}] = \mu$, $\mathrm{var}(\overline{X_n}) = \sigma^2/n$, which means that $\mathbb{E}[Y_n] = 0$ and $\mathrm{var}(Y_n) = 1$ for all $n \in \mathbb{N}$. For this reason (7.8) is often known as a *standardization* of $\overline{X_n}$. If $\overline{X}_n = \mu$ then $Y_n = 0$. More generally, how far away Y_n is from zero corresponds to how far $\overline{X_n}$ is away from μ.

Its difficult to underestimate the importance of the next result. It shows that if the X_n have finite variance then $Y_n \overset{d}{\to} N(0,1)$, where $N(0,1)$ denotes the standard normal distribution, regardless of the distribution of the i.i.d. random variables X_n. This is extraordinarily useful from an experimental point of view, because it tells you something that you should expect to see happen in *every* experiment – provided you can take repeated independent samples and they have finite variance. It allows statistical tests to be constructed without knowing the precise distribution of the random quantities involved, and that allows experimental science to quantify how likely a particular experiment (repeated sufficiently many times) is to have observed some important fact and not just random chance. In *uncertain* situations, the scientific basis for how well we understand the world around us comes primarily from the central limit theorem. We will discuss its history in Section 7.5.1.

Theorem 7.5.1 (Central Limit Theorem) *Let (X_n) be a sequence of i.i.d. random variables each having finite mean μ and finite variance σ^2. Let Y_n be given by (7.8). Then Y_n converges in distribution to the $N(0,1)$ distribution, as $n \to \infty$.*

Our proof will be based on Lemma 7.4.7 and the following lemma, which is a slight extension of the famous result that $\lim_{n\to\infty}(1+\frac{x}{n})^n \to e^x$.

Lemma 7.5.2 *Let $y \in \mathbb{R}$ and let α_n be a sequence of real or complex numbers such that* $\lim_n \alpha_n = 0$. *Then for all $y \in \mathbb{R}$, as $n \to \infty$ we have*

$$\left(1 + \frac{y + \alpha_n}{n}\right)^n \to e^y \tag{7.9}$$

The proof of Lemma 7.5.2 is an exercise in real analysis, which is included as Problem 7.11. We are now ready to prove the central limit theorem.

Proof of Theorem 7.5.1: Without loss of generality we assume that $\mu = 0$ and $\sigma = 1$. We can recover the general result from this special case by replacing X_n by $(X_n - \mu)/\sigma$. Hence $\mathbb{E}[X_1] = \mu = 0$ and $\mathbb{E}[X_1^2] = 1$. We have also that $\mathrm{var}(X_1)$ and $\mathbb{E}[X_1^1]$ are finite.

Our strategy is to show that the characteristic function of ϕ converges to the characteristic function of the $N(0,1)$ distribution, and then apply Theorem 7.4.5. Let ψ be the common characteristic function of the X_n, given by $\psi(u) = \mathbb{E}[e^{iuX_1}]$ for all $u \in \mathbb{R}$. Let ϕ_n be the characteristic function of Y_n for each $n \in \mathbb{N}$. Using independence and Lemma 7.4.3 we have that

$$\phi_n(u) = \mathbb{E}\left[e^{i\frac{u}{\sqrt{n}}(X_1+X_2+\dots+X_n)}\right] = \psi(u/\sqrt{n})^n. \tag{7.10}$$

Applying Lemma 7.4.7 to ψ, for all $y \in \mathbb{R}$ we have

$$\left|\psi(y) - \Big(1 + iy\mathbb{E}[X_1] - y^2\mathbb{E}[X_1^2]\Big)\right| \leq \mathbb{E}\left[\max\left\{|yX_1|^2, \frac{|yX_1|^3}{6}\right\}\right].$$

Setting $y = u/\sqrt{n}$, using that $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[X_1^2] = 1$,

$$\left|\psi(u/\sqrt{n}) - \left(1 - \frac{u^2}{2n}\right)\right| \leq \mathbb{E}\left[\max\left\{\frac{|X_1|^2}{n}, \frac{|X_1|^3}{6n^{3/2}}\right\}\right].$$

Let us write $\theta_n(u) = \mathbb{E}\left[\max\left\{\frac{|X_1|^2}{n}, \frac{|X_1|^3}{6n^{3/2}}\right\}\right]$. Putting the above into (7.10), we obtain

$$\left(1 - \frac{u^2/2}{n} - \theta_n(u)\right)^n \leq \phi_n(u) \leq \left(1 - \frac{u^2/2}{n} + \theta_n(u)\right)^n$$

and thus

$$\left(1 + \frac{-u^2/2 - n\theta_n(u)}{n}\right)^n \leq \phi_n(u) \leq \left(1 + \frac{-u^2/2 + n\theta_n(u)}{n}\right)^n. \tag{7.11}$$

We have $n\theta_n(u) = \mathbb{E}\left[\max\left\{|X_1|^2, \frac{|X_1|^3}{6n^{2/2}}\right\}\right]$. Note that $\max\left\{|X_1|^2, \frac{|X_1|^3}{6n^{2/2}}\right\} \in \mathcal{L}^1$ by Lemma 4.6.1 because $|X_1|^2 \in \mathcal{L}^1$, and that $\max\left\{|X_1|^2, \frac{|X_1|^3}{6n^{2/2}}\right\} \to 0$ pointwise as $n \to \infty$. By the dominated convergence theorem $n\theta_n(u) \to 0$ as $n \to \infty$. From (7.11) and Lemma 7.5.2 we thus have $\phi_n(u) \to e^{-\frac{u^2}{2}}$ for all $u \in \mathbb{R}$. In Exercise **7.9** we showed that $e^{-\frac{u^2}{2}}$ is the characteristic function of the $N(0,1)$ distribution. Hence from Theorem 7.4.5 we have $Y_n \xrightarrow{d} N(0,1)$, as required. ■

7.5.1 Further discussion ($\star$)

The central limit theorem is arguably the single most important result in probability and statistics. It is likely to be one of the earliest things you learned about, even though we could not give a rigorous proof until now.

The picture we know today was pieced together gradually by many different people. The sheer number of mathematicians that were involved in efforts to prove central limit theorems, particularly in the late 19th and early 20th century, makes a concise description of its history all but impossible. The modern statement of Theorem 7.5.1 is not attributed to any single author. The term ‘central limit’ is generally thought to have been introduced by Hungarian mathematian George Pólya in 1920.

One of the first examples is the case of tosses of a fair coin (taking e.g. taking +1 for heads and −1 for tails), studied by de Moivre in 1733 and later extended by Laplace – see Exercise **7.10** for this case. Around 1890, Chebyshev was the first mathematician to consider formulating the central limit theorem in terms of a sequence of independent random variables. Before that point, the results were formulated in terms of the convergence of particular probabilities. As we noted at the start of Chapter 5, the foundation of probability theory in terms of Lebesgue measure was also not established at that time. In fact, early proofs of what became the central limit theorem were based on extremely difficult calculations. Substantial effort was made to find less convoluted proofs, eventually leading to the argument given in these notes.

The version given in Theorem 7.5.1 has been extensively generalised during the 20th century. For example, conditions are known under which the central limit theorem holds for dependent sequences of random variables, for martingales, for stochastic processes with independent increments, and for cases where the normalization differs from that of subtracting the mean and dividing by the standard deviation.

We will discuss only a few such results here. If the i.i.d. sequence (X_n) is such that $\mu = 0$ and $\mathbb{E}[|X_n|^3] = \rho^3 < \infty$, the *Berry-Esseen theorem* gives a useful bound for the difference between the cdf of the normalised sum and the cdf Φ of the standard normal. To be precise we have that for all $x \in \mathbb{R}, n \in \mathbb{N}$:

$$\left|\mathbb{P}\left(\frac{S_n}{\sigma\sqrt{n}} \le x\right) - \Phi(x)\right| \le C\frac{\rho}{\sqrt{n}\sigma^3},$$

where $C > 0$.

We can also relax the requirement that the sequence (X_n) be independent. Consider the *triangular array* $(X_{nk}, k = 1, \dots, n, n \in \mathbb{N})$ of random variables which we may list as follows:

$$\begin{array}{ccccc}
X_{11} & & & & \\
X_{21} & X_{22} & & & \\
X_{31} & X_{32} & X_{33} & & \\
\vdots & \vdots & \vdots & & \\
X_{n1} & X_{n2} & X_{n3} & \dots & X_{nn} \\
\vdots & \vdots & \vdots & \vdots & \vdots
\end{array}$$

We assume that each row comprises independent random variables, but we allow random variables within each row to have dependences. Assume further that $\mathbb{E}[X_{nk}] = 0$ and $\sigma_{nk}^2 = \mathbb{E}[X_{nk}^2] < \infty$ for all k, n. Define the row sums $S_n = X_{n1} + X_{n2} + \cdots + X_{nn}$ for all $n \in \mathbb{N}$ and define $\tau_n = \mathrm{var}(S_n) = \sum_{k=1}^n \sigma_{nk}^2$. *Lindeburgh's central limit theorem* states that if we have the asymptotic tail condition

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{1}{\tau_n^2} \int_{|X_{nk}| \ge \epsilon\tau_n} X_{nk}^2(\omega)\, d\mathbb{P}(\omega) = 0,$$

for all $\epsilon > 0$ then $\frac{S_n}{\tau_n}$ converges in distribution to a standard normal as $n \to \infty$.

The highlights of this chapter have been the proofs of the law of large numbers and central limit theorem. There is a third result that is often grouped together with the other two as one of the key results about sums of i.i.d. random variables. It is called the *law of the iterated logarithm* and it gives bounds on the fluctuations of S_n for an i.i.d sequence with $\mu = 0$ and $\sigma = 1$. The result is quite remarkable. It states that almost surely,

$$\liminf_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log(n)}} = -1, \qquad \limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log(n)}} = 1.$$

This means that (with probability one) if $c > 1$ then only finitely many of the events $S_n > c\sqrt{2n\log\log(n)}$ occur but if $c < 1$ then infinitely many of such events occur. Similarly, the other way up, at -1. This gives a *very* precise description of the long-term behaviour of S_n.

7.6 Exercises on Chapter 7

On the Borel-Cantelli lemmas

7.1 Let $k \in \mathbb{N}$. Prove that in a sequence of independent coin tosses, infinitely many runs of k consecutive heads will occur.

7.2 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let (A_n) be a sequence of events.

(a) Show that $\{A_n \text{ e.v.}\} \subseteq \{A_n \text{ i.o.}\}$.

(b) Show that $\{A_n \text{ i.o.}\}^c = \{A_n^c \text{ e.v.}\}$ and deduce that $\mathbb{P}[A_n \text{ i.o.}] = 1 - \mathbb{P}[A_n^c \text{ e.v.}]$.

(c) Show that

$$\mathbb{P}[A_n \text{ e.v.}] \;\leq\; \liminf_{n\to\infty} \mathbb{P}[A_n] \;\leq\; \limsup_{n\to\infty} \mathbb{P}[A_n] \;\leq\; \mathbb{P}\left[A_n \text{ i.o.}\right].$$

On convergence of random variables and laws of large numbers

7.3 Let (X_n) be a sequence of i.i.d. random variables such that $\mathbb{P}[X_n = 1] = \mathbb{P}[X_n = 0] = \frac{1}{2}$. Show that $X_n \stackrel{d}{\to} X_1$ as $n \to \infty$, but that this convergence does not hold in probability.

7.4 (a) Let (X_n) be a sequence of random variables such that $\mathbb{P}[X_n = n] = \frac{1}{n^2}$ and $\mathbb{P}[X_n = 0] = 1 - \frac{1}{n^2}$. Show that $X_n \stackrel{a.s}{\to} 0$ and $X_n \stackrel{L^1}{\to} 0$.

(b) Let (X_n) be a sequence of independent random variables such that $\mathbb{P}[X_n = n] = \frac{1}{n}$ and $\mathbb{P}[X_n = 0] = 1 - \frac{1}{n}$. Show that X_n does not converge to zero almost surely or in L^1.

(c) Let (X_n) be a sequence of random variables such that $\mathbb{P}[X_n = n^2] = \frac{1}{n^2}$ and $\mathbb{P}[X_n = 0] = 1 - \frac{1}{n^2}$. Show that $X_n \stackrel{a.s.}{\to} 0$, and that X_n does not converge to zero in L^1.

(d) Let (X_n) be a sequence of independent random variables such that $\mathbb{P}[X_n = \sqrt{n}] = \frac{1}{n}$ and $\mathbb{P}[X_n = 0] = 1 - \frac{1}{n}$. Show that $X_n \stackrel{L^1}{\to} 0$, and that X_n does not converge to almost surely to zero.

(e) Deduce that $X_n \stackrel{\mathbb{P}}{\to} 0$ in all of the above cases.

7.5 Show that the following sequence (X_n) and candidate limit X of random variables converges in probability but not almost surely.

Take $\Omega = [0,1], \mathcal{F} = \mathcal{B}([0,1])$ and $\mathbb{P}$ to be Lebesgue measure. Take $X = 0$ and define $X_n = \mathbb{1}_{A_n}$ where $A_1 = [0,1/2], A_2 = [1/2,1], A_3 = [0,1/4], A_4 = [1/4,1/2], A_5 = [1/2,3/4], A_6 = [3/4,1], A_7 = [0,1/8], A_8 = [1/8,1/4]$ etc.

7.6 Let (X_n) be a sequence of random variables, and let X and Y be random variables.

(a) Show that if $X_n \stackrel{d}{\to} X$ and $X_n \stackrel{d}{\to} Y$ then X and Y have the same distribution function i.e. $F_X = F_Y$.

Hint: If two right-continuous functions are equal almost everywhere, they are equal everywhere.

(b) Show that if $X_n \stackrel{\mathbb{P}}{\to} X$ and $X_n \stackrel{\mathbb{P}}{\to} Y$ then $X = Y$ almost surely.

7.7 Examine the proof of the weak law of large numbers (Theorem 7.3.1) carefully. Show that the conclusion continues to hold if the requirement that the random variables (X_n) are i.i.d. is replaced by the weaker condition that they are identically distributed and *uncorrelated*, that is $\mathbb{E}[X_m X_n] = \mathbb{E}[X_m]\mathbb{E}[X_n]$ whenever $m \neq n$.

7.8 (a) Let X be a random variable and suppose that $X \geq 0$. Show that for any $a \in (0,1]$ we have $\mathbb{E}[\min(1,X)] \leq a + \mathbb{P}[X \geq a]$.

(b) Let (X_n) be a sequence of random variables. Suppose that $X_n \geq 0$ for all $n \in \mathbb{N}$. Show that as $n \to \infty$,

$$X_n \xrightarrow{\mathbb{P}} 0 \quad \text{if and only if} \quad \mathbb{E}[\min(1, X_n)] \to 0.$$

Hint: Recall Markov's inequality (Lemma 4.2.3).

On characteristic functions and central limit theorem (Δ)

7.9 (a) Let X be a random variable with the $N(0,1)$ distribution. Let ϕ be the characteristic function of X. Show that $\phi'(u) = -u\phi_Y(u)$ and hence show that $\phi(x) = e^{-x^2/2}$.

Hint: Use the result of Exercise 4.17 to show that the real and imaginary parts of $y \to \phi(u)$ are differentiable.

(b) Extend part (a) to cover $X \sim N(\mu, \sigma^2)$.

7.10 The first central limit theorem to be established was due to de Moivre and Laplace. In this case each X_n takes only two values, $\mathbb{P}[X_n = 1] = p$ and $\mathbb{P}[X_n = -1] = 1-p$, where $p \in [0,1]$. Write down the theorem in this special case, in terms of $S_n = X_1 + X_2 + \cdots + X_n$, and explain how it can be used to justify binomial approximations to the normal distribution.

7.11 The following result, which you do not need to prove, comes from real analysis.

Dini's Theorem. *Let $a < b$. For each $n \in \mathbb{N}$ let $f_n : [a,b] \to \mathbb{R}$ be a continuous function and suppose that $f_n \leq f_{n+1}$ for all n. If the pointwise limit $f_n \to f$ exists, and if $f : [a,b] \to \mathbb{R}$ is continuous, then $f_n \to f$ uniformly.*

Let $f_n(x) = (1 + \frac{x}{n})^n$. Use the AM-GM inequality from Exercise 6.4 to show that $f_n(x) \leq f_{n+1}(x)$ for $x \geq 0$ and all $n \in \mathbb{N}$. Hence, use Dini's theorem to prove Lemma 7.5.2.

Challenge questions

7.12 (a) Let (X_n) be a sequence of random variables and let $c \in \mathbb{R}$ be deterministic. Suppose that $X_n \xrightarrow{d} c$. Show that $X_n \xrightarrow{\mathbb{P}} c$.

(b) Let (X_n) be a sequence of independent random variables and suppose that $X_n \xrightarrow{\mathbb{P}} X$. Show that there exists deterministic $c \in \mathbb{R}$ such that $\mathbb{P}[X = c] = 1$.

7.13 A sequence (X_n) of random variables is said to convergence *completely* to the random variable X if

$$\sum_n \mathbb{P}[|X_n - X| \geq \epsilon] < \infty \qquad \text{for all } \epsilon > 0.$$

(a) Show that for sequences of independent random variables, complete convergence is equivalent to almost sure convergence.

(b) Find a sequence of (dependent) random variables (X_n) that converges almost surely but not completely.

Appendix A

Advice for revision/exams

There are two different exam papers, one for MAS31002 and one for MAS61022. For both exams the rubric reads

> *Candidates should attempt ALL questions. The maximum marks for the various parts of the questions are indicated. The paper will be marked out of 50.*

Within these notes, material marked with a (δ) is examinable only for MAS61022, and is non-examinable for MAS31002. Material marked with a ($\star$) is non-examinable for everyone.

- You will be asked to solve problems based on the material in these notes. There will be a broad range of difficulty amongst the questions. Some will be variations of questions in the assignments/notes, others will also try to test your ingenuity.
- You may be asked to state important definitions and results (e.g. more than one past exam has asked for definition of a measure).
- You will not be expected to reproduce long proofs from memory. You are expected to have followed the techniques within the proofs when they are present, and to be able to use these techniques in your own problem solving.
- There are marks for attempting a suitable method, and for justifying rigorous mathematical deductions, as well as for reaching a correct conclusion.
- If you apply an important result that has a name e.g. 'the Dominated Convergence Theorem' you should mention that name, or something similar e.g. 'by dominated convergence' or 'by the DCT'.

Revision activities

The most important activities:

1. Check and mark your solutions to assignment questions.
2. Learn the key definitions, results, and examples.
3. Do the past exam papers, and mark your own solutions.

Other very helpful activities:

4. Work through, and check your solutions, to non-challenge questions in the notes.

Of course, you should have been working on these questions throughout the year, which is why they are lower priority now. You do *not* need to look at the challenge questions as part of your revision – these are intended only to offer a serious, time consuming challenge to strong students.

In all cases, you are welcome to come and discuss any questions/comments/typos. Please use office hours or email to arrange a convenient time.

Appendix B

Solutions to exercises

Solutions are omitted from the printed lecture notes. You can find them inside the online version of these notes.